Phonological Reduction and Intelligibility in Task-Oriented Dialogue

Catherine Frances Sotillo

Thesis submitted for the degree of Doctor of Philosophy
University of Edinburgh
1997

Declaration

I hereby declare that I composed this thesis myself. The work described is my own research with the exception of the intelligibility experiments presented in Chapter 6, which were run as part of the research activity of the Dialogue Working Group at the ESRC funded Human Communication Research Centre†. My personal contribution to the collaborative work of this group was as follows:

- All maps used in the HCRC Map Task Corpus were created by myself, utilising drawings produced by Dr R. C. Shillcock;
- Some of the original transcription was done by myself, along with several cycles of checking;
- The referential coding of first and second mentions in all 128 dialogues was undertaken by myself, with checking by Dr E. G. Bard;
- The basic design of the intelligibility experiments was devised by Dr E. G. Bard; the detailed design, the selection of material, the excerpting of speech stimuli and the creating of the experimental tapes was done by myself;
- The experiments were run in the Department of Psychology at the University of Glasgow by Dr A. H. Anderson and Dr G. Doherty-Sneddon, with occasional assistance from myself;
- The raw response data was converted by myself into response matrices suitable for input to statistical analysis;
- Statistical analyses were run by both Dr E. G. Bard and myself;
- Interpretation of the results was based on collaborative discussion between Dr E. G. Bard, Dr A. H. Anderson and myself.

†The financial support of the Economic and Social Research Council UK (ESRC) is gratefully acknowledged.

Catherine F. Sotillo
Edinburgh
3rd October 1997

Acknowledgements

Writers of theses with long gestation periods find, when they look back over the last several years, that they have a veritable army of people to thank for their help and support. To those of you I have forgotten, forgive me: it is now getting hard to remember those early days when I was wrestling with ideas that were going nowhere slowly. To the rest of you, whether you are mentioned individually or not, thank you. You know who you are, and this thesis would never have got to the stage it is at now without you.

I am particularly indebted to the following people for their various contributions:

- Eddie Rooney and Michael Broe for their advice and support during my demoralising attempt to find acoustic evidence for nasal assimilation;
- Matthew Aylett for providing additional data and ANOVAs, and for his stimulating and thought provoking tea-time discussions;
- Laurence Molloy for his useful tutorial on the meaning of *k*-score duration;
- all the subjects who leant of their ears, whether they were judging nasal quality, accentedness, or trying to recognise words;
- my fellow Dialoggers at the Human Communication Research Centre for the interest they have shown and the insightful comments they have offered: Anne Anderson, Jean Carletta, Gwyneth Doherty-Sneddon, Stephen Isard, Andrew Merrison, Alison Newlands, Henry Thompson;
- my second supervisor, Bob Ladd, and the members of my ever-varying troika for their guidance, enthusiasm, and gentle coercion to get on and finish;
- Jim Miller and Ronnie Cann for believing I would finish;
- Ethel Jack for not letting me forget;
- Jan McAllister for her constant encouragement;
- my office mates, past and present, especially Jacqueline Kowtko and Bethan Davies for providing the final carrots;
- the HCRC Computer Support Team, who have always been swift to respond to my queries (Order of Merit goes to Lex Holt for solving so many LaTeX queries); also Laurence Molloy, Mike Morony, and anyone else who has written awk or perl scripts for me;
- Louise Kelly and Possidonia Gontijo, for sympathising as I struggled to come to terms with CELEX;
- my wonderful mother, who travelled 800 miles just to cook, clean and launder while her daughter exchanged domestic responsibility for a computer terminal;

- Georgina Chapman for reminding me that the only way to go is forwards;
- Mike, Jennifer and Alan for always being there when needed;
- Carla, Hazel, Lesley-Ann, Lynn and everyone else at the Blacket Teddies, young and old, who has looked after and entertained Helen while I have been writing up;
- Helen, for helping me sort out my priorities in life;
- and John, for the help, patience, tolerance and understanding he has shown for the last seven years. The Moon? Pah! The *Sun* and back!

There remain two people to thank. First, my father, who sadly will never read this, but whose playful wit and sheer joy in using the English language first kindled my interest in matters linguistic, and prompted a young architecture student to take Linguistics I "just for the hell of it".

Finally, if there is one person of whom it can appropriately be said "without whom not" it is my supervisor, Ellen Bard, who clearly has the patience of Job combined with the wisdom of Solomon. In knowing when to encourage, when to humour, when to cajole, and when finally to bark, Ellen has no equal. Her enthusiasm throughout this long campaign has been my lifeline. The successful completion of this thesis is undoubtedly due to Ellen's dogged determination that it should – and her touching faith that it could – be done. Ellen, I salute you!

For

Gerald Antony Leonard Bennett

1931 – 1984

Abstract

This thesis explores the implications of Lindblom's theory of Hyper- and Hypo-articulation (Lindblom, 1983b, 1990a) for word intelligibility and the likely application of phonological reduction processes in spontaneous discourse, using data from the HCRC Map Task Corpus.

Lindblom claims that variability in articulatory clarity is a reflection of speakers' assessments of their listeners' informational requirements: speakers hyper-articulate when listeners require maximum acoustic information, and economise on articulatory effort when listeners can supplement the acoustic input with information from other sources. To prevent speakers from over-economising to a point of unintelligibility, hypo-articulation is governed by a constraint of lexical distinctiveness: speakers hypo-articulate only while listeners are able to distinguish the target from competing lexical items.

Three main questions are addressed. First, do the informational needs of the listener affect the articulatory clarity of words produced in spontaneous conversation? A series of intelligibility experiments shows that repeated mentions of landmark names are less intelligible than their introductory mentions, independent of which speaker utters either mention, and who can see the landmark on their map. Although the results can be interpreted as supporting Lindblom's view, textual Givenness (Prince, 1981) is shown to depend upon what the speaker knows, rather than what the speaker believes her listener to know. The reduction in clarity associated with an increase in available information is not necessarily as listener-oriented as the H & H theory proposes.

Second, do phonological reduction processes such as word-final /d/-deletion or place assimilation contribute to intelligibility loss? Although reduction processes are found to be more prevalent in tokens from spontaneous discourse than in matched citation forms, they generally fail to account for effects of repetition. An increase in assimilation is found for repeated mentions of nasal-final stimuli in pre-velar position, but no effect is found for assimilation in pre-labial position, or for word-final /d/-deletion, nor is an effect found for the duration of schwa in metrically Weak initial syllables of polysyllabic words.

Third, does lexical competition predict the likelihood of targets undergoing reduction processes? Error responses from the intelligibility experiments are used to define lexical competition in terms of 'loose' cohort sets. The application of three reduction processes is shown to alter the set of lexical competitors for some but not all target words. However, the presence/absence of lexical competitors appears to have no effect on observed levels of reduction: for example, speakers assimilate first mentions of landmarks whether or not the assimilation results in an acoustic output that activates similar lexical competitors. I conclude that Lindblom's distinctiveness constraint is ill-supported: speakers hypo-articulate beyond the point at which lexical distinctiveness can be maintained.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis explores the consequences of variation in speech production on the successful recognition of words. It takes as a primitive the fact that speech is used to communicate a message to a listener; the goal is therefore one of message understanding rather than phonetic transcription. What matters is recognising the words rather than the sounds that compose them.

The thesis focuses on Lindblom (1990a)'s theory about speech production – the theory of Hyper- and Hypo-articulation, or H & H theory – and assesses the claims it makes in the light of intelligibility differences found for repeated coreferential mention in spontaneous dialogue. It looks for evidence to support a predicted relation between reduction in articulatory clarity (hypo-articulation) and repetition by investigating the contribution made by three phonological reduction processes to intelligibility loss, using words excerpted from conversations about the location of landmarks on a map.

Speakers are shown to be less sensitive to their listeners' informational needs than Lindblom's theory suggests: rather than maintain an accurate model of their listeners' discourse, speakers rely on generalising from their own discourse experience. Although in many situations these two views of discourse will coincide, they do not always. That speakers' articulatory effort is more self- than listener-oriented is further demonstrated by evidence that speakers hypo-articulate beyond the point of lexical distinctiveness, even when introducing information which is New to the discourse.

1.1 Issues addressed by thesis

The H & H theory proposes that speakers are required to place in the acoustic stream only that information which is unavailable to the listener from other sources. In other words speakers hyper- or hypo-articulate according to their listeners' informational needs. In essence, the H & H theory combines the principle of economy of effort with a listener-oriented constraint of distinctiveness: speakers articulate with the minimum of effort required to maintain sufficient distinctiveness for the listener.

While this may be a perfectly plausible view, I argue that the theory as it stands is incomplete. I highlight three areas of concern, each of which is addressed by the empirical work of this thesis.

First, I suggest that the source of data marshalled in support of Lindblom's position is not the most appropriate. The H & H theory makes predictions about speech production in a communicative context and yet the data on which it is based derive from research on controlled laboratory speech. I argue that an appropriate test of the theory's claims requires an analysis of speech produced in natural conversation where both speaker and listener are engaged in the exchange of information.

Second, I indicate that the H & H theory lacks any detailed account of the basis on which speakers might assess their listeners' informational requirements. The Given/New distinction (Prince, 1981) is shown to provide a linguistically relevant and robust division of information structure that can be used to test the claims of the H & H theory: the H & H theory predicts that speakers may hyper-articulate New information but will hypo-articulate what is already Given.

Third, I consider the implications of the H & H theory's distinctiveness constraint. Essentially, the distinctiveness constraint is presented as a requirement for listeners to be able to distinguish between *words*. I interpret this to mean that speakers should refrain from hypo-articulation when it introduces lexical ambiguity. The distinctiveness constraint can therefore be tested by finding examples of phonological reduction processes that result in an increase in lexical competition. The H & H theory predicts that in introducing New information, speakers should refrain from employing phonological reductions when their application leads to lexical ambiguity.

1.2 Approach of thesis

The H & H theory predicts that a speaker's articulatory behaviour is conditioned by the informational needs of her[1] listener. It has two core components, and, in association, two key participants: the principle of economy of effort involves a speaker, and the notion of sufficient perceptual distinctiveness concerns a listener. In introducing the latter component Lindblom implies that speaking behaviour has a communicative goal, that there is an intention on the speaker's part to convey information to a listener – information which the listener must decode from the speech signal. However, the data Lindblom uses to argue in favour of his H & H theory is derived from investigations of carefully recorded scripted text and CVC syllables – the standard tools of the laboratory phonetician's trade. The problem with such data is that they lack both a communicative goal and a genuine addressee. I suggest that a more appropriate source of material would be natural non-scripted dialogue with a clear requirement for the successful exchange of information. One of the aims of this thesis, therefore, is to test the validity of Lindblom's view with respect to data from a corpus of spontaneous goal-directed dialogues: the HCRC Map Task Corpus. In such dialogues there is a genuine requirement for listeners to understand and act upon the information conveyed by the speaker. If the H & H theory is correct, then we ought to find evidence of speaker economy in just such a corpus of speech, where speakers will be assessing the needs of their listeners and adjusting articulatory effort accordingly.

If economy in articulatory effort is guided by what the speaker believes the listener already knows, then we need to consider what sources of knowledge are relevant for the listener, and how the speaker may model her listener's belief state. There is ample evidence in the literature to suggest that the more contextual information there is available, the more predictable a word becomes, and consequently the more redundant will be the acoustic information associated with a word's articulation. To assess how economical she can be in articulating any word, then, a speaker must have some notion of what information her listener requires to interpret the incoming acoustic signal successfully, and what information is already known.

Consequently, I elaborate on Lindblom's position by incorporating a division of information structure along the lines of Prince (1981)'s taxonomy of 'Given' and

[1] Here and throughout the thesis the speaker is presented as female and the listener as male. Similarly Map Task Instruction Givers are female, and Followers male. The gender distinction helps in the anaphoric resolution of potentially ambiguous pronouns.

'New', where 'Given' refers to something previously mentioned or known about in the discourse, and 'New' refers to something that is being introduced into the discourse for the first time. The Given/New distinction provides a robust framework for discussing the kinds of information requirements that are known to have consequences for speech production and perception. Within such a framework, it is hypothesised that speakers will economise in articulatory effort when referring to Given information, and articulate more carefully when introducing New information. Subsequent reference to a Given entity should not require the same articulatory effort since some information about the referent is already available to the listener.

Previous work by Fowler and Housum (1987) and others (Lieberman, 1963; Hunnicut, 1985) has demonstrated a relation between the Given/New distinction and intelligibility: words which refer to Given information are shorter and less intelligible than words which refer to New information, a relation referred to as the "Repetition Effect". The ease or difficulty with which a group of listeners can recognise a word token when it is excised from context is a reflection of the articulatory effort or economy employed by a speaker in producing the token. Intelligibility, then, is one way of assessing the articulatory information a speaker has made available to the listener. A series of intelligibility experiments is described which focuses on the nature and content of the discourse model that a speaker builds of her listener's knowledge state. The experiments explore the extent to which speakers are genuinely cooperative: does a speaker refrain from hypo-articulation when she receives feedback from her listener that he has no visible access to the referent, for example?

Speakers are shown to be less cooperative than the H & H theory predicts. The results indicate that the discourse model which a speaker constructs for her listener is broad rather than fine-tuned, and is egocentrically based on the speaker's own knowledge of the discourse. Speakers reduce the intelligibility of repeated mentions regardless of who has visible access to the referent. Intelligibility is also lost when speakers introduce landmarks on the second occasion they encounter a map, even though the referent is New for the listener.

Having established an effect of repetition on intelligibility I ask to what extent intelligibility loss can be accounted for by the application of phonological reduction processes such as place assimilation and segment deletion. Two studies explore the effects of word-final processes, while one examines the effects of a reduction process at the start of a word. In all cases the question at issue is

whether reference to Given entities is characterised by the more frequent application of phonological reduction rules *vis-à-vis* control productions in the form of citation forms. The first study explores the relation between repetition and place of articulation assimilation in word-final alveolar nasals. The second study involves a phonetic analysis of /d/-deletion, comparing the duration of word-final stop consonants for introductory and repeated mentions of Map Task landmark names. The final study analyses the incidence of schwa syncope in the word-initial syllables of metrically Weak-Strong polysyllables.

Although a repetition effect is established for the assimilation of /n/ preceding velars, no effect of repetition is found for pre-labial assimilation, nor for /d/-deletion or schwa syncope.

In two of the studies, economy in articulatory effort involves a reduction in duration, with segment deletion characterising the application of maximum economy; in the case of assimilation, articulatory economy changes the identity of a segment. Both segment deletion and change in segment identity may result in the activation of an alternative lexical hypothesis, depending on the level of competition at that point in the word. This fact allows us to test the validity of Lindblom's claim that hypo-articulation is constrained by lexical distinctiveness. The requirement on a listener is to recognise the words uttered by the speaker, rather than the phonemes *per se*. That is, a speaker ought to be able to economise articulatory effort so long as her listener can successfully recognise the words that compose the speaker's message.

Because of the phonotactic constraints of a language, not all segments compete with all others at all times. Given the structure of the lexicon and the nature of lexical competition, it is not always necessary to recognise every segment in order to recognise the whole word. If, as Lindblom proposes, the natural tendency of speakers is to economise on articulatory effort, that is, to articulate with the least effort required for understanding, then presumably speakers need articulate clearly only those segments which function to distinguish meaning. In other words, speakers will control articulatory clarity according to which parts of a word listeners cannot afford to miss.

I review the literature on word recognition, focusing on two current models: the Cohort theory (Marslen-Wilson and Tyler, 1980; Marslen-Wilson, 1987; Marslen-Wilson and Gaskell, 1992) and TRACE (McClelland and Elman, 1986). Two important notions are introduced: lexical competition, and the Uniqueness Point of a word – the point at which a word diverges from its same-sounding competitors

and becomes lexically unique. I propose that a speaker's assessment of information redundancy ought to include information about the structure and content of the mental lexicon. That is, in considering what information her listener requires her to place in the acoustic signal, a speaker will take account of the possible lexical confusions that might arise from a lack of articulatory clarity, or "*in*sufficient distinctiveness". It is hypothesised that where confusion with another segment would activate a competing word candidate, or when another word candidate offers a close phonetic competitor for the target phoneme, speakers cannot afford to hypo-articulate.

An analysis of the assimilation in first mentions of landmark names shows no difference between words that do and do not have close competitors when assimilated.

1.3 Organisation of thesis

The chapters and their contents are as follows:

Chapter Two This chapter raises the problem of speech variability and its treatment and introduces Lindblom's explanation for the lack of acoustic invariance: the theory of Hypo- and Hyper-articulation. The chapter highlights three main problems with the theory: the data on which the theory is based; the failure of the theory to provide an explicit account of the sources of signal-independent information; and the implications of the distinctiveness constraint. The chapter concludes by introducing three phonological reduction processes which can be used to test whether lexical distinctiveness is maintained: word-final place assimilation, stop-deletion and pre-stress schwa syncope.

Chapter Three Here I present an account of the process of word recognition, introduce the key concepts of competition, frequency and uniqueness, and describe two models which currently dominate the literature. I also detail the results of gating and priming studies which have investigated the effects of phonological mismatch on word recognition.

Chapter Four This chapter provides the detailed account of information structure in discourse that the H & H theory requires if it is to have any kind of predictive power. I introduce the Given/New distinction and discuss the cognitive

representation of Givenness in terms of referent accessibility. I also discuss Clark (1992)'s view of discourse as essentially 'cooperative'.

Chapter Five Here I describe the HCRC Map Task Corpus which supplied the data used in the empirical work discussed in the subsequent chapters. I present an account of the corpus design, and provide illustrative examples of the text and speech produced. I include a description of the segmentation criteria adopted in the speech analyses.

Chapter Six In this chapter I present the results of a series of intelligibility experiments run in conjunction with fellow researchers at the Human Communication Research Centre. By manipulating what the speaker and listener could and could not see, we were able to explore the perceptual consequences of changing the Given/New status of landmark referents in the HCRC Map Task, and the degree to which speakers' articulation showed sensitivity to the informational needs of their listener.

Chapter Seven This chapter explores the effects of word-final place assimilation on intelligibility. Two assessments of assimilation are undertaken: an acoustic analysis based on pole-zero decomposition, and a perceptual judgement task, requiring expert phoneticians to judge how [m]-like, [n]-like and [ŋ]-like each nasal sounds.

Chapter Eight I describe an investigation of word-final /d/-deletion, using observations and measurements taken from spectrographic and time/amplitude waveform displays. The significance of the location of phonological reduction within a word is raised, and the effect of repetition on the duration of word-final stops compared with that of stressed vowels.

Chapter Nine The issue of where in the word reduction occurs is explored further. I analyse the effects of schwa reduction in polysyllabic words with Weak-Strong (WS) word-initial syllables. I also test for a repetition effect for stressed vowel duration over the materials as a whole.

Chapter Ten Here I explore the relation between lexical competition and phonological reduction. I test the standard Cohort-based definition of competitor

by seeing how well it accounts for the incorrect responses offered as alternative candidates by the subjects in Chapter 6's intelligibility experiments. Having established a definition of lexical competition based on these analyses, I consider the possible relation between lexical competition and the likelihood of application of phonological processes of segment deletion and assimilation.

Chapter Eleven The final chapter summarises the main research findings and discusses their implications for Lindblom's H & H theory. Suggestions are made on how the work could be extended.

Chapter 2

Explaining speech variability: Lindblom's theory of Hyper- and Hypo-articulation

2.1 Introduction

The core of the empirical work undertaken in this thesis concerns the effect of phonological reduction processes on the intelligibility of words excerpted from non-scripted conversational speech. By implication, this work is concerned with the **variability** with which tokens of a word may be produced and the problem such variability may present to word recognition.

Consequently this chapter starts with a discussion of speech variability and *why* it is problematic; I introduce Lindblom (1990a)'s theory of Hyper- and Hypo-articulation – the H & H theory – as representing a much cited 'explanation' for the lack of invariance in the speech signal. The H & H theory holds that a speaker economises articulatory effort according to the amount of signal-independent information available to her listener to help him decode the acoustic input. I present the evidence Lindblom offers in support of his view, and then level two main criticisms at the theory. The implications of these two criticisms are explored more fully in Chapters 3 and 4. The chapter concludes with a discussion of the kind of data required for an appropriate test of Lindblom's theory. I identify three phonological reduction processes which characterise hypo-articulation in connected speech, and which it can be shown have potentially harmful consequences for the process of word recognition. It is argued that such processes present a challenge to an important component of Lindblom's H & H theory: the distinctiveness constraint.

2.2 Speech variability

One phenomenon that distinguishes spoken from written language is the variability of individual tokens of any one word form. In written forms of language, most especially in type-script form, tokens rarely differ from one instance to another. The use of UPPER CASE, and font changes such as *italic* and **bold** introduce only limited variation: while italic differs from bold, all italicised tokens are identical to each other, for example. The spoken word, on the other hand, is extremely variable: no two spoken word tokens are identical, even if uttered by the same speaker within the same utterance. Indeed, it has been stated that:

> *"If there is one finding which can be said to most universally characterise speech processes, it is the finding of variability."*
> *(Lindblom* et al., *1986, page ii)*

So, then, in what ways do spoken tokens of the same word vary from one production to another? What are the possible sources of variation? The difficulty experienced by automatic speech recognition systems serves to illustrate the ubiquitous nature of variation in speech production. The solutions employed to make the problem of automatic recognition a tractable one reflect our inability to cope with many of the known sources of variance. According to Klatt:

> *"The cumulative effects of this variability are so great that current systems designed to recognise only the isolated digits zero to nine have considerable difficulty doing so in a speaker-independent manner."*
> *(1986:304)*

Essentially, speech recognition systems – until recently at least – tended to opt either for restrictions on the number of speakers to be recognised, or on vocabulary size. Different speakers vary in the way they produce tokens of a word; speaker A's token of word Y may sound more like speaker B's token of word Z. Consequently, restricting the number of speakers a system has to cope with will reduce the chances of recognition failure. Similarly, a small vocabulary – or **lexicon** – constrains the competition between different vocabulary items, so that despite variation in an item's pronunciation, it may not sound sufficiently like any other word in the lexicon for confusion to arise. Both these constraints can be implemented whether the automatic recognition device is being applied to isolated words or connected speech. Indeed, the restriction to recognising isolated words represents a further constraint on possible sources of variability by removing the

effects of connected speech processes, that is, the contextual effects that occur at linguistic boundaries.

In other words, one approach to automatic recognition is to eliminate certain sources of variability by restricting the type of input allowed. We could think of this as a reduction in the level of noise that speech variation introduces: it is an attempt to cut down rather than solve the problem.

However, as we shall see later, it can be argued that some sources of articulatory variability are in fact regular and predictable, so that variability need not always be viewed as noise at all, but as an additional source of information. Clearly, the more we understand about how and when articulation varies, the better our automatic systems will be at coping with less restricted, that is, more natural input.

Two sources of variability that cause problems for the successful recognition of speech have been identified above: differences between speakers, and differences in the way a word form is articulated which result in its sounding like another similar word. It is primarily the second of these issues that is addressed by the work in this thesis, but given the nature of my data it will be necessary at least to acknowledge the problem of variance across speakers. The next two sections therefore consider speaker and word token variation respectively.

2.2.1 Speaker variation

The empirical work of this thesis involves analyses of the production and recognition of isolated word tokens excerpted from natural spontaneous dialogue. Intelligibility studies require tokens of large numbers of different word forms to avoid asking subjects to recognise the same word form more than once. As a result, word tokens were excerpted from a corpus of 128 separate dialogues between a total of 64 different speakers. Comparing tokens across dialogues necessarily means comparing tokens from one speaker with those of another. It is important therefore to be aware of the variation in speech production between different speakers.

Sources of speaker differences have traditionally been divided into **structural**, i.e. anatomical (or organic) differences – such as variation in the shape and size of the vocal apparatus – and **functional** differences, which concern the idiosyncratic manner of an individual's speech (see, for example, Garvin and Ladefoged, 1963; Wolf, 1972), although it has been argued (Nolan, 1983) that the

organic/functional dichotomy is a misleading oversimplification.

Structural differences include the length and shape of the vocal tract, both of which play a significant role in determining the resonant qualities of an individual's speech, while functional dimensions along which speakers vary include habitual speech rate and amplitude.

It is precisely because the acoustic output can vary so greatly from one speaker to another that speech technology has been tempted to extract speaker-specific cues which are sufficiently stable and invariant, as the basis for the automatic recognition of speaker identity.

To avoid the problems associated with speaker differences, therefore, analyses undertaken in this thesis compare individual tokens from unscripted dialogues with tokens of the same word form produced by the same speaker in citation (list-reading form), rather than comparing tokens directly across speaker.

2.2.2 Token variation: effects of connected speech processes

It has been well established in the literature that phonological reduction processes are prevalent in continuous speech (see Shockey, 1974; Brown, 1977; Shockey and Bond, 1980; Dalby, 1984, among others). Rather than being stable in form, words in connected speech are subject to variation in duration, amplitude, and spectral composition. The factors which govern such variation are many and varied, and include a word's position within an utterance (Oller, 1973; van Santen and Olive, 1990), the location of prosodic boundaries (Rakerd *et al.*, 1987; Wightman *et al.*, 1992), speech style (Shockey, 1974; Fowler, 1988; McAllister, 1989), speech rate (Lindblom, 1963; Dalby, 1984; van Son and Pols, 1989), and, as we shall see later, predictability (Lieberman, 1963; Hunnicut, 1985; Shields and Balota, 1991). Whilst some of these factors, such as phrase boundary location, result in segment/word lengthening (Cooper and Danly, 1981), our main concern here is with *reduction* processes, that is, changes to duration, amplitude, and/or spectral composition which result in **attenuated** tokens. This is because it is the **lack** of acoustic information which has significant consequences for lexical access, most specifically, for matching the acoustic input onto more than one lexical representation.

Variation can occur on any segment: all segments can be stretched or compressed, produced at varying amplitudes, or modified in terms of their spectral characteris-

tics. Some segments more readily undergo changes than others, either for reasons associated with the articulatory processes that produce them (continuants – fricatives and nasals, for example – are, by definition, easier to manipulate in terms of duration, than are non-continuants such as stop consonants), or for reasons to do with 'distinctiveness', in other words, the level of competition between similar sounding segments (see Hawkins and Warren 1994 and, also, the discussion in Section 2.3.3 below).

Both vowel and consonant segments are subject to reduction processes. Research on **vowel** reduction has tended to concentrate on modifications to spectral characteristics, in particular plotting changes in the relations between F1 and F2 in the formant 'vowel space' (Tiffany, 1959; Lindblom, 1963; Öhman, 1966; Gay, 1978; van Bergem and Koopmans-van Beinum, 1989; Bates, 1995). Changes to the spectral characteristics of **consonants** arise from the application of a large number of different processes such as glottalisation, palatalisation, flapping, gemination and deletion.[1]

As McAllister *et al.* (1990:1) observe, "the consequence of this degree of variability at the segmental level is an enormous range of potential pronunciations for individual words". Kohler (1990), for example, offers, for the inflected German definite article *dem*, the following set of pronunciations which, he suggests, reflect a **reduction hierarchy** of progressively stronger reductions:

(2.1) [deːm] ⟶ [dem] ⟶ [dɐ̯m] ⟶ [dəm] ⟶ [dm] ⟶ [bm] ⟶ [m]

The changes from fullest form [deːm] to maximally reduced [m] encompass a variety of processes including shortening, centralisation, weakening, deletion and assimilation.

A necessary prerequisite for the application of processes like these is an appropriately conditioning **phonological context**. A segment will only undergo palatalisation, for example, in the presence of a neighbouring palatal or palato-alveolar consonant. However, the presence of a phonological context which conditions a particular reduction does not, of itself, guarantee the application of the reduction 'rule'. Few, if any, connected speech processes are compulsory. Indeed, it is the **optionality** of application which marks these processes out as *phonological* rather than phonetic in nature (see Cohn, 1993).

Some of the factors likely to influence whether or not a segment undergoes a

[1] For a catalogue of the possible processes that segments can undergo see, for example, Brown (1977) or Gimson (1980).

reduction process are reasonably well-known and understood; these include lexical, prosodic and syntactic conditions, as well as more 'global' characteristics like speech rate and style. For example, Cooper and Paccia-Cooper (1980) showed that palatalisation and alveolar flapping are more likely to be suspended when the segments precede major syntactic boundaries or verb deletion sites, and in words bearing emphatic stress. The incidence of palatalisation and flapping was also found to relate to word frequency: low frequency words are less likely to undergo reduction. Subsequent work (Cooper *et al.*, 1983) explored the effects of speech rate and showed that palatalisation across word boundaries was more frequent among characteristically fast speakers and at fast rates of speech than elsewhere. Dalby (1984) found similar effects of speech rate on the incidence of schwa-deletion in both pre- and post-stress position. Non-linguistic factors such as speaker familiarity have also been shown to affect the application of reduction processes. Adult speech to children is more reduced than speech by adults to other adults (Shockey and Bond, 1980; Bard and Anderson, 1983), while shorter word tokens occur in speaking to friends than in speech to strangers (McAllister, Sotillo, and Bard, 1991).

Nevertheless, our ability to predict the precise level of reduction for a particular token is far from perfect. One of the aims of this thesis, therefore, is to extend our knowledge about the likelihood of words being reduced, based on the predictions made by Lindblom's H & H theory.

2.2.3 Why variability is a problem

Clearly variability in production is not a problem for users of language. Although there are certain characteristics of the speech signal which make the recognition of words less than straight-forward to model, as speakers and listeners we experience few apparent difficulties. It is in attempting to model the recognition process itself that the difficulties arise. The problem of variability centres on the issue of mapping input to representation: how do we get from variable acoustic patterns to the invariant phonological representation of a word? In other words, the invariance problem concerns the nature of linguistic representation, and, specifically, the relation between phonetics and phonology.

In orthographic as well as linguistic representation, claims Lindblom (1983b), we characterise utterances in terms of *sequences of discrete units* such as words, syllables, vowels and consonants; and yet, examination of the speech signals themselves highlights the fact that:

> *"there seems to be no unique set of acoustic properties that will always be present in the production of a given unit (feature, phoneme, syllable* etc.*) and that will be reliably found in all conceivable contexts."*
> *(Lindblom, 1983b, page 156)*

Lindblom refers to this as the **invariance issue**. The invariance issue assumes that language is essentially quantal[2] in nature, that it is featurally and segmentally structured, but that this quantal structure is not immediately evident in the speech signal. Thus, there is a fundamental incompatibility between the phonologist's view of language as a finite, hierarchical set of discrete linguistic units and the raw material of the phonetician's research: an infinite set of overlapping acoustic and articulatory elements.

Lindblom and MacNeilage (1986) offer two alternative solutions to the problem. The first is to assume that *"all the information required for perception is in the signal"* (1986:129). In this case the problem of invariance is that of identifying the correct signal attributes and applying the appropriate transformations. In other words, invariance is indeed in the signal; we just have not found it yet.

Lindblom, however, argues that any attempt to find an answer to the invariance issue in the signal itself is bound to fail. Work on the dynamic aspects of vowel articulation[3] leads him to observe that:

> *"since we know that successful communication is possible in spite of far-reaching and frequent reduction and omission of acoustic cues we have very strong reasons for doubting that the question of invariance is exclusively or mainly a measurement problem."*
> *(Lindblom, 1983b, page 156)*

Lindblom's preferred resolution of the invariance problem is to assume that *"all of, part of, or none of, the information required for perception is in the signal"* (Lindblom and MacNeilage, 1986, page 129). In this case, none of, the remaining part of, or all of the information required for perception is contributed by the **current internal state of the listener's perceptual system**. In other words,

> *"Talkers realise phonetic segments in production and phonetic structure is specified in the acoustic signal, only in so far as explicit signal information is needed to supplement implicit contextual listener*

[2]By 'quantal' I mean that there are *minimal units* of linguistic structure, such as phonemes, or distinctive features, and that there is a (presumably universal) set of speech sounds and/or phonetic dimensions that is *finite*. According to Lindblom and Engstrand (1989) the impression of finiteness may be an illusion, based on a descriptive need to 'quantise' phonetic sound shapes into a manageably large set of phonetic symbols. Under such a view languages are quantal at the phonological level, but only *weakly* so at the phonetic level.

[3]see Section 2.3.2.3

'knowledge'."
(Lindblom and MacNeilage, 1986, page 130)

According to this view invariance is not necessarily a physical phonetic phenomenon, and ultimately it can be defined only at the level of listener 'comprehension'.

Similarly, Kiparsky (1986) suggests that we should ask whether the patterning of speech variation is "an irreducible statistical component of linguistic knowledge", or is the result of speakers

> *"striving ... to optimise their speech according to ... [such] kinds of functional considerations ... [as] simplifying their articulation, making their output maximally comprehensible, and taking account of whatever social values they discover are attached to the various options – goals which may be weighted in different ways depending on the circumstances."*
> *(Kiparsky, 1986, page 423)*

This second approach to the issue of invariance *refocuses* the problem: it concentrates on its origin and purpose, and presents phonetic variability as a core component of speech production rather than problematic 'noise' to be explained away. Variation is viewed as an inherent, rule-governed, functional property of language which, at the phonetic level, is "an information bearing aspect of speech, rather than information-burying noise" (Dalby, 1984, page 12).

If this is the case, then what are the rules that govern such variation? How can we predict the likelihood of a token undergoing, for example, word-final place assimilation, or palatalisation?

In the following section I detail a theory of speech production which has evolved from the view that invariance is not to be found in the acoustic signal: Lindblom's theory of Hyper- and Hypo-articulation, or, the H & H theory. As we shall see, Lindblom takes the view that variability in the signal is associated with the speaker's beliefs about the listener's requirements for acoustic information, which vary according to what other information is available to the listener from the context of utterance. It is this variability in her listener's informational requirements, claims Lindblom, that governs the likelihood of a speaker producing reduced tokens.

2.3 The theory of Hyper- and Hypo-articulation

Lindblom (1990a) offers a theory of speech production that explains speech variability in terms of **speaker choice** in relation to **listener need**: speakers specify phonetic structure in the acoustic signal to the extent that it is needed by the listener to supplement signal-independent contextual knowledge. In this section I outline my interpretation of Lindblom's position. I then go on to discuss the problems that arise from this view, some significant omissions, and the sort of data that could be used to test the theory.

2.3.1 A biological theory of language

Before describing the core elements of Lindblom's H & H theory, a short diversion is offered by way of situating Lindblom's basic philosophical stance.

Underlying much of Lindblom's work is an implicit desire for a **biological** theory of language that aims at deriving linguistic elements and processes deductively rather than postulating them axiomatically. Lindblom claims to adopt a **functionalist** position: the constraints of speaking, listening, and learning "interact in complex ways to delimit humanly possible sound patterns" (Lindblom, 1983a, page 217). He contrasts this with nonfunctionalists (Chomsky, for example) for whom the primitives of linguistic theory are abstract and formal, rather than substantively (biologically and socially) motivated. In particular, he states that:

> *"whereas [the functionalist approach] derives the fundamental units and process of linguistic structure deductively from independent premises anchored in psychological and physical realities, [the nonfunctionalist position] postulates them axiomatically for formal reasons."*
> *(Lindblom, 1983a, page 218)*

Lindblom sees this latter view as circular, describing linguistic data in terms of units postulated specifically to describe the data. He argues that while this may be descriptively adequate, such an approach fails to achieve explanatory status, in that it cannot account for *why* the data patterns one way rather than another. It cannot answer the question of where the linguistic units describing the data come from (see Lindblom *et al.*, 1983; Lindblom and Engstrand, 1989).

Evidence of Lindblom's general philosophical stance can be found in a review of Kelso *et al.* (1986)'s paper on Event Perception and Action Theory (EPACT). Lindblom and MacNeilage (1986) describe Kelso, Saltzman and Tuller's work

(hereafter KST) as "a healthy reaction against too much mentalism and cognitive theorising about language" (Lindblom and MacNeilage, 1986, page 119), and a rejection of the 'black boxes' of information processing frequently postulated by computer models. Lindblom aligns himself with the **ecological** perspective of Gibson, Johansson and Bernstein. In their critique, Lindblom and MacNeilage go on to state that "Though purporting to be a contribution to an ecological theory of speech, the KST approach to articulatory modelling leaves the listener and the communicative and social functions of speech conspicuously out of the picture" (1986:126). Lindblom's own H & H theory is presented quite decidedly as an ecological position, sketching a scenario in which the listener, the environment, and all the functions that human speech subserves, play a part.

Lindblom's desire to find an **external** source of explanation for speech variability leads him to consider the characteristics of **motor behaviours** in general. The following section details how he applies the principle of 'Economy of Effort' to linguistic phenomena.

2.3.2 The economy of effort principle

When Lindblom explores the idea of pronounceability, considering the physical limits of speech production, he observes that in normal speech the production system is rarely driven to its limits. There is, instead, an under-exploitation of potential capacity.

> *"The frequent occurrence of reduction processes in speech, here typified by vowel observations, provides [...] evidence that extremes are avoided. Extreme displacements and extreme velocities are avoided." (Lindblom, 1983a, page 231)*

How and why might this be? Lindblom seeks an explanation in the characteristics of general motor behaviour, demonstrating that the same attributes are evident in speech behaviour. He suggests that motor behaviour is characterised by the following three attributes:

- **output-oriented control**: the *goal* of the movement determines the sequence of motor events
- **plasticity**: actions can be *adjusted according to the demands of the situation*

- **economy**: the manoeuvre should *expend no more physiological energy than necessary*

In other words, motor control is **teleologically** organised, i.e. purpose-driven (Granit, 1977). These key notions can also be found in speech, claims Lindblom.

2.3.2.1 Output-oriented control

Lindblom argues that speech is similar to other motor behaviours in that muscular force levels are tailored to the needs of the situation: the sequence of motor events is determined by the goal of the movement. For speech the situation is necessarily the communicative context in which the articulation takes place. Speech is *purposeful*: a speaker produces an utterance in order to communicate some intention to a listener, whether the intention is to convey propositional information, emotional state or whatever. For the goal to be achieved, i.e. the successful communication of intention, the listener must be able to recognise the string of incoming acoustic elements and decode it into meaningful chunks in order to ascertain the speaker's intended meaning.

Speech motor behaviour ought therefore to be tailored to the needs of the listener. Levelt refers to "the canonical ecological context of talking: the speaker's participation in conversation" (1989:2). That is, 'normal' speech behaviour consists in conversation between two or more participants. And if conversation is a collaborative activity, it can be successful only if the speaker "respects, or takes into account, the rights, capabilities, propensities, and feelings of the other parties" (Levelt, 1989:65)[4].

If Lindblom is correct in his assumption that speech mirrors motor behaviours in general then we would expect to find that articulatory behaviour is influenced by listeners' requirements.

2.3.2.2 Plasticity

The notion of plasticity is essential to Lindblom's position with respect to the invariance problem; variability in the speech signal, far from being linguistically irrelevant, is in fact a product of speaker *adaptation*: "phonetic gestures and signals are modulated and tuned adaptively in accordance with on-line communicative and socio-linguistic demands" (Lindblom *et al.*, 1992, page 357). In other

[4]See, also, the discussion in Sections 4.3 and 4.9 below

words, non-invariance arises from the demands of the communicative situation. Speakers illustrate plasticity in their ability to adjust pronunciation style, and also in studies of **compensatory articulation**.

For example, Lindblom, Lubker, and Gay (1979) examined formant frequency data for four Swedish vowels [i, u, o, ɑ], produced by six speakers in two conditions: uttered with spontaneous jaw position (i.e. 'normal') versus with the position of the lower jaw controlled by means of a **bite block**. Vowels were produced in isolation, with a total of 18 repetitions per vowel for the normal condition. Nine repetitions of each vowel were produced before the bite block sessions, with the remaining repetitions occurring after the bite block vowels. For the bite block condition subjects were told that the experiment might be "somewhat more difficult", but that they should try to pronounce the vowels so that they sound as similar as possible to the normal vowels, and that they should make them sound alike in as few attempts as possible. This enabled the researchers to compare the very first bite block productions – which reflect subjects' immediate unpractised responses to an entirely novel (or at least highly unfamiliar) compensatory motor task – with subsequent productions, which would demonstrate any learning effects. Lindblom *et al.*'s principle finding was that despite the physiologically unnatural jaw opening involved with bite blocks, all six subjects were able to produce formant frequency patterns the majority of which fell within the ranges of variation observed for the normal vowels. What is more, no practice was required by the subjects in order to achieve this. The researchers hypothesise that the 'instantaneous' learning of such a 'novel task'[5] is possible:

> *"neither because speakers draw extensively upon past similar experience nor because special motor mechanisms distinct from those of natural speech are invoked but primarily because normal speech motor programming is indeed 'compensatory'. In other words, it operates in a context-sensitive mode to achieve listener-oriented goals*[6] *and since 'contexts' constitute a practically infinite class of events the programming has to be 'creative', that is it must be capable of handling conditions never experienced before."*
> *(Lindblom* et al., *1979, page 147)*

In other words, subjects appear to be able to achieve **acoustic equivalence** despite having to compensate for the fixed jaw position, because of the inherent

[5] Actually, this situation is not as novel as Lindblom *et al.* imply. People are extremely used to talking with their mouths full, and also compensating for wearing orthodontic braces, dental bridges, EPG palates *etc.*

[6] It should be noted that there were no listeners directly involved in the experiment (other than the experimenters themselves).

adaptability of motor behaviours in general.

2.3.2.3 Economy

Lindblom offers two examples of articulatory simplification which illustrate the principle of motor economy in speech: consonant-vowel coarticulation, and vowel reduction. He argues that these two phenomena suggest two physiologically based conditions: "a *synergy constraint* governing static spatial relations among articulators, and a *rate constraint* operating dynamically on articulatory movements" (Lindblom, 1983a, page 220). Phonological assimilation rules, argues Lindblom, exemplify language adaptations to both synergy and rate constraints.

2.3.2.3.1 Coarticulation The motor events of any sequence of phonemes overlap in space and time. As Lindblom observes:

> *"the signal cannot be unambiguously segmented into temporally nonoverlapping chunks corresponding to linear sequences of phonemes, syllables and words"*
> *(Lindblom, 1983b, page 156)*

The resulting spatial and temporal overlap of adjacent gestures explains (in part) why we find variability in the speech signal, despite the invariant nature of the intended articulatory configuration or underlying target. This general phenomenon of overlap is called **coarticulation.**

For example, from spectrograms of [d] articulations in the context of [ʏdʏ], [ɑdɑ] and [udu] frames, Öhman (1966) observed that the value of F2 at the boundary of the [d] segment appeared to correlate with the location of F2 at the vowel steady state. If the articulatory configurations underlying the acoustic facts are considered, it can be seen that, while the point of contact of the tongue tip remains invariant for [d], the tongue body contour varies, bearing a strong resemblance to its shape in the adjacent vowel. In Lindblom's terms the tongue-tip gesture is coarticulated with the tongue-body position for the vowel environment (Lindblom, 1983a, page 221).

A numerical model of coarticulation was developed by Lindblom and Sundberg (1971) to explore the natural degrees of freedom of the articulatory system. Using various input parameters for lips, mandible, tongue body, tongue blade (i.e. tip), and larynx, in conjunction with a set of geometric rules, the model generates an articulatory profile with sufficient information to allow acoustic calculations to be

made. The model can generate all possible combinations of parameter values that are compatible with the production of a particular 'segment' at a specified place of articulation, such as a tongue-tip stop consonant. Comparison with X-ray data on VCV utterances reveals that the model is capable of providing realistic descriptions of real articulatory profiles. However, the model *over-generates*: in human speech, extreme values of the parameters are avoided. That is, "normal speech seems to exploit no more than a fraction of the degrees of freedom that are in principle available for articulation" (Lindblom, 1983a, page224). That extremes in motor behaviour are avoided should come as little surprise to Lindblom. Regularly pushing the body to physiological limits tends to lead to physical damage, as evidenced by the injuries sustained by professional dancers, athletes *etc.*

The discrepancy between possible and actual articulatory configurations is interpreted by Lindblom as reflecting a **synergy** constraint on tip-body coordination. Speakers avoid extreme values, producing instead, gestures which are coarticulated with neighbouring segments. The extent to which any two gestures are coarticulated will depend on the need to avoid extreme displacements, in other words, the need for economy:

> *"The degree of coarticulation is manifested in the extent to which the vowel environment is allowed to colour the formant frequency pattern of the consonant and to influence the tongue-body shape underlying tip closure. Evidently the degree of coarticulation is related to the severity of the rule saying that extreme displacements are avoided" (Lindblom, 1983a, pages 225-226)*

Since the presence and degree of coarticulation is variable, Lindblom proposes that coarticulation must be a result of "motor control optimisation processes", which – whenever other contingencies permit – contribute toward making speech gestures *more economical.*

2.3.2.3.2 Vowel reduction Whilst coarticulation arises from a synergy constraint on the coordination of articulators, vowel reduction is interpreted by Lindblom as an illustration of a **rate** constraint on articulation.

It was observed above (see Section 2.5.3) that vowel reduction has traditionally been associated with 'centralisation'. However, Stevens and House (1963) attributed their findings of observed vowel undershoot to *two* articulatory processes: centralisation and **contextual assimilation** (what I refer to above as 'coarticulation'). An investigation of Swedish vowels pronounced under varying timing

conditions and in systematically varied consonantal environments was undertaken by Lindblom (1963) to evaluate the roles of centralisation and coarticulation in the process of vowel reduction. He combined the eight short, or lax, Swedish vowels: [ɪ, e, ʏ, æ, a, ɵ, ɔ, ʊ] with three consonantal contexts: [bVb], [dVd], and [gVg], and embedded these CVC syllables in four carrier phrases which varied rhythm (phrasal stress) and word order. Thus CVC syllables were produced with varying stress patterns. The talker was asked to project the rhythmic pattern of the carrier phrases onto a basic periodic beat played through headphones, in order to assure a constant speaking rate. In a supplement to the main analysis, the talker was asked to produce stressed tokens of the CVC syllables in isolation, with speech rate being varied by altering the timing of the periodic signal in steps from 0.5 to 6.0 cps. In this way Lindblom was able to compare reduction arising from stress variation with that from change in tempo.

Lindblom found that unstressed vowels and vowels produced at faster speaking rates behaved in much the same way: they were *shorter* in duration and *more reduced* in quality (defined in terms of formant values) compared with their stressed or slow speech rate counterparts. Lindblom concludes that it is "immaterial whether a given length of the vowel is produced chiefly by the tempo or the degree of stress. Duration seems to be the main determinant of reduction." (1963:1780).

In his 1963 paper, Lindblom proposes, therefore, a model of **duration-dependent undershoot**, whereby formant frequencies are less likely to attain their ideal target values as vowel length decreases. Vowel targets are defined explicitly in terms of the asymptotic values of their first two formant frequencies and are found to be independent of consonantal context and duration. In other words, targets are an invariant attribute of the vowel. Lindblom suggests that a speaker's intention is always the same: she aims at producing a full, maximally distinct vowel. However, realisation of the intended vowel is affected by the temporal overlap in the timing of motor commands to the articulators. If commands occur in close temporal succession "the system may be responding to several signals simultaneously and the result is coarticulation" (1963:1778). This model of vowel reduction excludes an independent and explicit centralisation process. While undershoot *may* result in more central formant frequencies this is by no means a logical necessity. Centralisation simply arises from contextual assimilation to neighbouring segments which themselves have a more central locus of articulation, such as might occur when a CVC syllable is embedded in a [hə–#] frame[7], or is preceded or followed

[7]as used by Stevens and House (1963)

by silence (when articulators will incline towards their rest position).

Subsequent research has shown that a model of vowel reduction based on duration alone is too simplistic. Undershoot is *not* an inevitable consequence of short duration (Kuehn and Moll, 1976; Gay, 1978). For example, Kuehn and Moll used cineradiography to record movements of various points on the tongue, lips and mandible of five speakers, looking at the effects on articulatory velocity of phonetic context and variation in speaking rate. They measured the displacement of each marker point and the duration interval for each transitional movement in a VCVC syllable, constructed from a set of three vowels [i, ɑ, u] and six consonants [p, t, k, f, s, l], and compared these with steady state vowel values derived from a [hV] context frame. In general, the farther a given speaker moved a primary articulator, the faster he generally moved it.

> *"At normal speaking rate velocity of movement is contingent on magnitude of displacement, which depends on phonetic context within speakers, and size of oral structures between speakers"*
> *(Kuehn and Moll, 1976, page 318)*

The effect of increased speaking rate was variable: for one speaker it always resulted in an increase in velocity; for two speakers there was usually an increase; for the remaining two speakers there was generally a decrease. Kuehn and Moll conclude that at a rapid speaking rate transition time is decreased, but "speakers have the option of either increasing velocity of movement or decreasing articulatory displacement" (1976:318). In other words, undershoot can be, and sometimes is, avoided by making more rapid approaches to targets.

Speaker variability is also exhibited in work by Moon and Lindblom (1989) in which they compare the amount of under- or overarticulation in **citation form** and **clear speech**, in an attempt to identify numerical rules that define reduction and elaboration. Five speakers of American English read lists of words in two conditions: at a rate and loudness spontaneously chosen by the subject (citation-form condition) and then with an explicit instruction to "overarticulate and to speak as clearly as possible as when communicating with a non-native listener" (1989:121) (clear speech condition). Words were chosen which involved front vowels in a [wVl] context: *wheel, will, well, wail,* in order to gain maximally sensitive acoustic indications of articulatory undershoot (the [wVl] frame should result in large F2 transitions). Vowel duration was systematically varied by embedding the monosyllabic [wVl] word in bi- and trisyllabic frames generated by appending *-ing* and *-ingby* or *-ingham* (place name endings), giving, for example: *will, willing,*

Willingham; *well, welling, Wellingby.* Repetitions of each vowel in a [hVd] frame were also recorded for each subject to provide a null context control. Each item was repeated five times.

Formant frequency was plotted as a function of vowel duration to investigate undershoot. All five speakers demonstrated vowel undershoot: vowels in monosyllables were less reduced than vowels in polysyllabic words (where the vowels are *shorter*). In other words, the shorter the vowel segment, the smaller the F2 excursion. The degree of target undershoot varied according to vowel, subject and speaking condition: tense vowels showed greater resistance to undershoot than lax vowels; while some speakers showed strong undershoot effects, others showed less (perhaps reflecting different interpretations of the instructions); in clear speech, speakers modified their pronunciation, using forms that were less reduced and more similar to the patterns observed in the relatively context-free environment of [hVd].

An exponential curve of the form depicted in (2.2) was fitted to the formant-versus-duration plots and showed that – for a given speaking style – the degree of vowel reduction can be predicted quantitatively from two factors: the size of the formant transition ($F_{ni} - F_{nt}$) and T, the duration of the vowel (Lindblom *et al.*, 1992, page 361).

(2.2)

$$F_{20} = k(F_{2i} - F_{2t})e^{-aT} + F_{2t}$$

In (2.2) above, F_{20} is F2 at the maximum of the formant excursion, k and a are constants, F_{2i} is the starting value of the transition, F_{2t} is the underlying vowel 'target', and T represents vowel duration[8]. This equation essentially replicates Lindblom (1963)'s earlier duration-dependent undershoot model. However, Lindblom, Brownlee, Davis, and Moon found that it was necessary to use different target values for the two different speech styles. Values of k, a and F_{2t} (the vowel target) had to be adjusted for the model to fit both the citation-form and clear speech measurements for any given talker and vowel. A higher target value was needed for the clear speech measurements, which had the effect of counteracting undershoot. The need for such adjustments suggests that speakers are capable of controlling the precise degree of reduction. The strong version of

[8] e^{-x} denotes a sigmoid function.

the duration-dependent undershoot model (Lindblom, 1963) must therefore be abandoned: vowel reduction is more than simply durationally induced contextual assimilation.

Using findings like these to support his view, Lindblom likens the speech production process to that of a mechanical system of mass, damping and spring components (see Lindblom, 1983a). In order to avoid undershoot, an increase in force is required to produce the desired displacement when duration is reduced. In other words, undershoot can be avoided by making more rapid approaches to targets, but this necessarily requires more force or *effort* on the speaker's part.

On what basis might a speaker choose to employ more or less effort? This is where Lindblom appeals to the output-oriented control of motor behaviour: speech has a purpose which is ultimately listener-oriented. Speakers articulate utterances in order to convey a message to their listener.

2.3.3 The distinctiveness constraint

Taken to its extreme the phonetic motor economy constraint would result in no articulation at all, or – assuming a need to convey some sort of message – an articulation that was so 'economical' to produce that it was totally unintelligible. Consequently, Lindblom introduces a teleological component to his theory: economising occurs only insofar as it is **purposeful**. The ultimate purpose of articulating is to convey meaning to a listener, therefore speakers can economise only so long as listeners are still able to recognise the message.

Lindblom assumes an *active* model of listening in which stimulus driven (bottom-up) and hypothesis-driven (top-down) processes interact i.e. the signal is interpreted by (subconsciously) applying linguistic redundancy rules and conceptual knowledge:

> *"It is by means of this 'predictive' strategy that listeners are able to restore, or compensate for, missing or degraded signal information and to perceive physically identical stimuli in different ways depending on the context. Accordingly, if we assign an important role to top-down processes in normal speech perception we evidently have a way of explaining why there should be no unique set of acoustic properties always present in the production of a given phoneme and that will reliably be found in all conceivable contexts."*
> *(Lindblom, 1983b, pages 156-157)*

According to Lindblom (1983b), whether two semantically distinct utterances will

be *perceptually* distinct is a function of:

- the properties of the acoustic signals
- their auditory representations
- the redundancy or predictability structure of the messages
- the listener's active use of such predictability

That is, we have explicit (speaker-generated) phonetic information and implicit (listener-generated) conceptual and linguistic information[9].

> *"Suppose that the talker is capable of making a (gross) predictive, running estimate of the implicit, listener-generated contribution. In the subconscious planning of an utterance the speaker can elaborate (overarticulate), or simplify (underarticulate) his articulatory gestures in accordance with that estimate."*
> *(Lindblom, 1983b, page 157)*

This view is expanded in Lindblom's sketch of his H & H theory. The H & H argument is based on four observations (Lindblom, 1990a, page 404):

- speech perception involves discrimination among items stored in the listener's lexicon. Lexical access is thus a function of the *distinctiveness* (rather than invariance) of the acoustic stimulus;
- lexical access is facilitated by signal-complementary processes i.e. 'knowledge';
- speech motor control is future-oriented i.e. purpose-driven;
- as the output constraints on a movement become less severe, it tends to default to some low-cost form of behaviour (system-oriented).

From these observations Lindblom deduces that: "the amount of explicit signal information minimally required for successful lexical access will vary between and within utterances" (Lindblom, 1990a, page405). He continues:

[9]The question of precisely where this listener-generated information might come from is raised in Section 2.4.3 and discussed more fully in Chapter 4.

> *"In the ideal case, the speaker estimates the running contribution that signal-complementary processes will make during the course of an utterance, and dynamically tunes the production of its elements to the short-term demands for either output-oriented control (hyper-speech) or system-oriented control (hypo-speech). What he/she needs to control is – not that linguistic units are actualised in terms of physical invariants (higher-order or whatever) – but that their signal attributes possess sufficient contrast, that is discriminable power that is sufficient for lexical access."* (sic)
> *(Lindblom, 1990a, page 405)*

In other words, there is a balance to be found between the demands of the listener for sufficiently recognisable input (output-oriented hyper-speech) and a natural tendency of the internal system towards economy of effort (system-oriented hypo-speech). For a speaker to find that balance she must assess the needs of her listener, economising articulatory effort up to but not beyond the point where sufficient contrast is maintained for successful recognition. Contrast, or distinctiveness, is the key to whether a speaker can afford to reduce clarity and under-articulate.

Thus Lindblom introduces a *distinctiveness constraint* which is essentially in conflict with the forces which bring about assimilations, reductions, syncope, and the like. Lindblom likens the process of speech production to a "continual tug-of-war between demands on the output on the one hand and system-based constraints on the other" (Lindblom, 1990a, page 420), in other words, an articulatory-perceptual cost-benefit trade-off.

Lindblom summarises the implications of the H & H theory for the invariance issue:

> *"the H & H theory assumes that, in all instances, speech perception is the product of both signal-driven and signal-independent information, that the contribution made by the signal-independent processes show short-term fluctuations and that speakers adapt to those fluctuations. It says that – whether communicatively successful or not – adaptive behaviour is the reason for the alleged lack of invariance in the speech signal. Hence it predicts that the quest for signal-based definitions of invariance will continue to remain unsuccessful as a matter of principle."* (sic)
> *(Lindblom, 1990a, page 431)*

2.3.4 Summary

Lindblom favours a functional and ecologically based explanation for phonetic variation, sketching a scenario in which the listener, the environment and all the functions that human speech subserves, play a part. His sympathies appear to lie more with the Gibsonian ecological approach to psychology than with the formal linguists of the Structuralist tradition.

Central to his writing is a strong desire to avoid postulating formal linguistic units axiomatically, preferring to deduce them from independent premises. Thus the economy of effort principle that appears to hold for motor behaviour in general is appealed to as an explanation for the variability to be found in phonetic data. In this way Lindblom can account for phonetic facts via a linguistically independent phenomenon of economy. In doing so, he places emphasis on the *dynamic* nature of speech production processes, as constrained by the human vocal apparatus and its degrees of freedom.

The economy of effort principle states that a manoeuvre should expend no more physiological energy than necessary. To keep speakers from articulating an unintelligible slurred mush, a teleological component is needed. Lindblom therefore introduces the notion of purposeful economy: articulatory simplifications are constrained by listener-oriented demands. Thus economy is checked by perceptual discriminability. The distinctiveness condition is essentially in opposition to the principle of economy. Speakers can reduce articulatory effort only as far as they maintain sufficient perceptual difference for successful lexical access by the listener.

2.4 Problems with the H & H theory

Although on the surface Lindblom's position may not appear unattractive, his exposition leaves a number of questions unanswered. This section attempts to address some of the theory's omissions.

There are four issues which Lindblom fails to deal with. First, the use of the terms 'hyper-' and 'hypo-' implies some sort of articulatory 'norm' against which an individual production is compared. Lindblom does not make it clear what this norm is or ought to be. Secondly, Lindblom lays himself open to the criticism that he may be looking for effects in the wrong data. The H & H theory is concerned with speech as communication: a meaningful discourse between two or

more agents. But the research on which the theory is based derives from tokens of read speech, recorded in a laboratory with little or no communicative function. Thirdly, Lindblom offers little discussion on the notion of distinctiveness; there is no account of what the perceptual constraints on production are, and how they might operate. Finally, the H & H theory fails to make explicit either the kind of signal-independent information that might be available to the listener, or how such information might be used to aid recognition.

I conclude that a fuller account is required if Lindblom's H & H theory is to be developed into the empirically testable account of linguistic behaviour that Lindblom clearly desires (see Lindblom, 1990a, page 404). This fuller account is presented in the ensuing chapters.

2.4.1 Where along the H & H continuum is 'normal'?

The H & H theory proposes that signal-independent information which the communicative situation makes available permits a speaker to articulate a word token less carefully than she would were she producing it as an isolated word out of context. A speaker needs to place in the acoustic stream the information which her listener requires to be able to recognise the word successfully; because of the contextual information already available which helps make the word more predictable, speaking to a listener in context ought to result in a speaker being able to underarticulate, compared with a carefully produced citation form. The relation between speech context and articulatory economy is summarised in (2.3) below.

Moon and Lindblom (1989) reported a difference in the degree of vowel undershoot which initially looks like a contradiction of the relations proposed in (2.3) between articulatory clarity and the presence/absence of a listener. They report moderate undershoot for vowels produced in *clear* speech WITH an hypothesised listener, with most undershoot in words read from a list with NO listener.

(2.3) a. **Citation:** list-reading
→ NO listener
→ **no** contextual support
→ *hyper*-speech

b. **Spontaneous dialogue:** conversation
→ WITH listener
→ **high**-contextual support
→ *hypo*-speech

However, it can be argued that the hyper-articulation found in their clear speech condition is a special case of over-emphasising distinctions not usually made in normal running speech, of attempting to produce maximally clear, target-achieving tokens in order to give the inexperienced or impoverished listener *additional* information[10].

This, of course, brings in to question what counts as a 'normal' as opposed to hyper- or hypo-articulated production of a word token? If hyper-articulation is somehow *adding* information, then what is the nature of the form being added to? Is this some normative production, and if so, what is its nature? The terms 'over'/'under', and 'hypo'/'hyper' imply some kind of base-line. So, what exactly is this base-line? It is not clear from what Lindblom says, whether he considers the context free open syllable vowel articulations as some sort of 'norm', with all other productions necessarily being underarticulated in comparison, or whether the citation forms are 'norms' with clear speech and [hVd] productions as over-articulation, or, indeed, whether clear speech is somehow central, with less and more undershoot evident in [hVd] and citation form contexts respectively. It is possible that Lindblom has no base-line production in mind but uses the words 'hyper' and 'hypo' as **relative** terms only.

Part of the confusion no doubt arises from differences in interpreting what is meant by the terms 'citation form', 'clear' speech, *etc.* The citation forms used in Moon and Lindblom (1989) are rather different from the traditional, carefully articulated 'dictionary entry'-type forms to which the term is usually applied. What is the difference between a 'base-line' production and a traditional citation form, for example? In the following section, I try to clarify the seemingly contradictory position highlighted above by considering the range of speech 'modes' available to the empirical investigator. My aim will be to find a suitable data source for

[10]based on the assumption that the non-native listener lacks knowledge of the phonology, syntax and lexicon of the language being spoken

testing hypotheses about speech reduction processes based on the H & H trade-off between economy of effort and the informational needs of a listener.

2.4.2 Testing the H & H theory – what data?

This section raises the issue of what constitutes a suitable source of data for testing predictions that arise from an H & H based view of speech production. Lindblom's research concentrates on the acoustic and articulatory facts in relation to traditional syllabic structures such as CVC frames, or nonsense words. But although Lindblom's analyses of reduction are based on laboratory recordings of read speech, he acknowledges the need to view speech as produced "not only in the laboratory but also in its natural, ecological settings" (Lindblom, 1990a, page 418). Clearly, in 'natural' contexts the amount of signal-independent information available to the listener is different from that available in standard laboratory conditions. Indeed, I believe it is hard to argue for a theory about the role of speech perception on production, based on data derived from experiments where there are no listeners to do any perceiving[11].

The likelihood of application of certain reduction processes varies according to speech mode. The selection of the appropriate speech material is therefore of paramount importance. I consider, below, the difference between isolated word production and tokens from connected speech; the advantage of spontaneous over read speech, and the number of participants that need to be involved in a *communicative* act.

2.4.2.1 Isolated words versus connected speech

Phonological reduction processes can, of course, only be studied for segments occurring in suitable environments. In some cases the required phonological environment is prosodic: schwa-deletion, for example, depends upon a particular patterning of Weak and Strong syllables[12] (Dalby, 1984). Other processes, such as glottalisation and assimilation, will only apply in a particular segmental environment. Thus, in order to examine the effects of certain reduction processes on word intelligibility, for example, it is necessary to consider the local phonological context in which the word occurs. An investigation into the occurrence of word

[11] I am assuming here that the presence of an experimenter is not equivalent to the existence of a conversation participant to whom a speaker is attempting to convey some piece of information.

[12] where 'Strong' refers to syllables with full vowels, and 'Weak' to syllables with reduced vowels (usually but not exclusively schwa) (Cutler and Norris, 1988)

boundary assimilation and stop deletion requires that the target word be embedded in material that is at least two words long, in order to allow for the effects of the conditioning environment at the boundary between the two words. All these – as well as a higher level linguistic environment – are found in connected, meaningful speech.

Carefully articulated list readings provide useful **citation forms** against which the connected speech tokens can be compared. This view of citation form is somewhat different from Moon and Lindblom (1989)'s definition. In Moon and Lindblom's experiments, citation forms were derived from speakers reading lists of words at a rate and loudness spontaneously chosen by the speaker, i.e. at a relaxed, 'normal' tempo. In the work described in subsequent chapters of this thesis I use the term 'citation form' to describe slow, carefully articulated tokens which will be used as a base-line measure.

2.4.2.2 Read versus spontaneous speech

Most of the speech material we encounter as listeners is uttered **spontaneously**; it is "conceived and composed by their speakers even as they are uttered" (Mehta and Cutler, 1988, page 136). In contrast the speech mode most often adopted in phonetic and psycholinguistic research is **read** speech: subjects read aloud material specially selected by the experimenter, ranging from lists of CVC syllables, through words within carrier phrases, to paragraphs of scripted text.

There is, of course, good justification for this. The need for tight control – or, better, the elimination – of variables irrelevant to the particular investigation being undertaken frequently requires very careful construction of materials. Clearly it is unlikely that such carefully designed utterances will 'walk in off the street' as spontaneous speech. Directing unscripted speech towards the production of specific words or phrases is difficult. One frequently adopted technique is to elicit descriptions of pictures or events depicted graphically. Alternatively, when eliciting dialogue, the experimenters may themselves participate in order to direct the topic of conversation towards a specific subject area (Hawkins and Warren, 1994).

A further advantage of using scripted material within a laboratory environment is recording quality: sound-proofed booths and state-of-the-art recording equipment result in high quality speech material that is amenable to manipulation via signal processing techniques. Spontaneous, unscripted speech material, on the other hand, has frequently been collected with poor quality portable equipment.

Subjects are more inclined to move and gesticulate when conversing rather than reading, which, again, can affect recording quality.

But can we be sure that findings established in the laboratory using scripted speech will generalise to speech that is unscripted (and vice versa)? The differences between read and spontaneous speech have been well documented (see Harris and Umeda, 1974; Mehta and Cutler, 1988; McAllister, 1989; Kowtko, 1996, amongst others). As well as prosodic differences between speech modes – read speech tends to be produced at a faster rate than spontaneous speech, which contains longer and more frequent pauses, hesitations and disfluencies – non-scripted natural speech is characterised by a greater degree of phonological reduction (Shockey, 1974) and syntactic simplification (Halliday, 1992). A study of hypo-articulation and the incidence of phonological reduction ought to concentrate on the speech mode that is characterised by articulatory economy, that is, non-scripted natural speech.

2.4.2.3 Monologue versus dialogue

While it is easier to control variables with read materials it is only when dealing with spontaneous speech that we can examine the effects on production and perception of a **genuine communicative context**, with the expression of ideas and information. If Lindblom is correct, and the presence of a **listener** has a significant influence on what and how something is said, then it is important to analyse data taken from spontaneous discourse involving both a speaker and a listener: i.e. dialogue.

In unscripted dialogue the two participants are engaged in a mutual exchange of information. Lindblom's H & H theory argues that the nature of the information previously exchanged may have some influence on the likelihood of a speaker attenuating articulatory clarity.

2.4.3 The role of signal-independent information

Frequent reference is made in the H & H theory to listeners' access to signal-independent information. It is not clear, however, what form this information takes, nor how it relates to speakers' articulatory behaviour. How, specifically, is the articulatory-perceptual cost-benefit criterion calculated, for example? What kind of record does a speaker maintain of her listener's access to and requirements for information? And, knowing this, how much economy can/will she employ?

Lindblom acknowledges that two utterances might have *identical* acoustic/perceptual properties and yet be recognised as having one of two different meanings according to the context: e.g. "*lesson five*" versus "*less than five*" (Lindblom, 1990a,b). If, as Lindblom argues, speakers only put in to the signal that information which they believe their listener cannot supply from elsewhere (i.e. from top-down hypothesis-driven knowledge sources) then a number of questions need to be addressed. For example, Lindblom suggests that a speaker makes a "**gross predictive running estimate**" of their listener's needs (Lindblom, 1983b). We need to ask:

- On what information will the speaker base these estimates?
- What form does this gross estimate take?
- *How* does a speaker compute what will be sufficiently distinct, given the state of her (gross) estimate?

In order to provide a workable version of the H & H model – to be able to apply the theory to the practice of predicting the degree of underarticulation in a given context – it will be necessary to pursue in greater detail how it is that a speaker might predict the informational requirements of her listener. In Chapter 4, therefore, I consider the kind of discourse model a speaker might generate in order to keep track of the information that has been mentioned, and explore what knowledge needs to be stored, and how it might be accessed. In particular I focus on the distinction between 'Given' and 'New' information and the linguistic means available to speakers to realise this distinction.

2.4.4 The problem with *sufficient* distinctiveness

The notion of 'sufficient distinctiveness' is problematic. What counts as sufficient, and at what level? It is not always clear from Lindblom's writing at what level of abstraction the contrast is necessary. Although he talks about **perceptual salience** as a relation that holds between different *phonetic* elements (Lindblom and Engstrand, 1989), in discussing the H & H theory, Lindblom clearly views **distinctiveness** as a characteristic of *words* (see Lindblom, 1990a, pages 404-405), tacitly acknowledging that if one adopts a goal-oriented approach to speech behaviour, what ultimately is being discriminated is at the semantic level of meaning, namely what is stored in the mental lexicon.

Lindblom does acknowledge that a theory of perceptual processing needs to take account of factors such as word frequency and neighbourhood size (Lindblom, 1990a, pages 409-412), which he sees as further examples of signal-independent considerations. The implication is that speakers can reduce articulatory effort so long as the distinctiveness of the *word* is maintained, where distinctiveness will be a function of how frequently the word appears in the language, of how like other words it is, and how frequent these other words are. In other words, articulatory reduction will be governed in part by issues of lexical competition.

One important phenomenon that Lindblom mentions in passing but does not pursue, is the issue of 'accidental gaps' in languages, i.e. cases that are compatible with phonotactic rules but are left lexically unexploited (see Lindblom, NacNeilage, and Studdert-Kennedy, 1983). These gaps are not predicted from motor constraints. They are just accidents, but they do have consequences for lexical access because they affect the size of the competitor set. Consider the minimal pairs in (2.4) and (2.5), where * indicates a non-word.

(2.4) a. *cap* – *gap*

b. *cat* – **gat*

(2.5) a. *rape* – **lape*

b. *rate* – *late*

c. *rake* – *lake*

Although /k/ and /g/ are distinctive in (2.4a), the gap in the lexicon exemplified in (2.4b) means that a speaker can reduce the VOT for *cat* without worrying that her listener will mis-recognise the target. Similarly, in (2.5a), the distinction between /l/ and /r/ is not lexically significant. If speakers are sensitive to lexical competition then they ought to be able to increase articulatory economy, i.e. hypo-articulate, by exploiting these sorts of gaps.

The question arises, therefore, whether lexical distinctiveness affects the degree of reduction a word token undergoes. Are word-final nasals less likely to be assimilated when the assimilated form leads to lexical ambiguity, for example? The lexical distinctiveness constraint appears to argue that speakers hypo-articulate up to but not beyond the point at which speakers can discriminate between words. If we can demonstrate that speakers do, in fact, hypo-articulate even when by doing so they introduce lexical ambiguity, then doubt must be cast on

the **lexicon** as the level at which 'sufficient contrast' is maintained. If speakers hypo-articulate *beyond* the level of lexical discriminability then successful recognition will need to depend not just on acoustic and lexical information but on a process of inferencing, based on pragmatic context.

To test whether speakers maintain lexical distinctiveness, three phonological reduction processes were selected that may make words which have undergone them more difficult to recognise: word-final place assimilation, word-final stop deletion, and the deletion of schwa in pre-stress position. The application of each of these three processes has the potential to introduce lexical ambiguity. For example, the assimilation of *been* in a pre-labial context to [bim] may activate the competitor *beam*; the deletion of word-final /d/ in *gold* may activate the competitor *goal*; the deletion of schwa in *collapse* may activate the competitor *claps*. The following section describes each of these processes in turn. Chapter 3 then considers the process of word recognition itself, and presents some of the key concepts essential to an understanding of how hypo-articulation can affect lexical processing.

2.5 A test for lexical distinctiveness: phonological reduction revisited

It was observed above that the application of certain processes of phonological reduction may have implications for the H & H theory's distinctiveness constraint. In the ensuing sections I describe three reduction processes which will be used in an empirical investigation into the relation between hypo-articulation and word intelligibility.

2.5.1 Assimilation

The phonological process of assimilation involves a phonological segment changing in some way to become more like its neighbour. This process may occur between segments within a word, or between segments across a word boundary. Abercrombie (1967) uses the term **juxtapositional assimilation** to differentiate between the word boundary process of assimilation and **similitude** which describes the regular coarticulatory accommodation of a segment to its phonetic context (such as the fronting of the articulation of [k] in *kit* compared with the [k] in *cat*). He defines juxtapositional assimilation as:

> *"changes in pronunciation which take place under certain circum-*

> *stances at the ends and the beginnings of words (changes at word 'boundaries', that is to say) when these words occur in connected speech, or in compounds ... [where] the final segment of the first and the initial segment of the second have become* similar *in certain respects in which they were different."*
> *(Abercrombie, 1967, pages 133-134)*

More recently, in line with autosegmental approaches, Nolan (1992) takes assimilation to be

> *"where two distinct underlying segments abut, and one "adopts" characteristics of the other to become more similar, or even identical, to it, as in cases such as [griːm peint]* green paint, *[rɛg kɑː]* red car, *[bæd̪ θɔːts]* bad thoughts."
> *(Nolan, 1992, page 262)*

Assimilation may take one of two forms. **Anticipatory** assimilation involves the final segment of the first word changing to become more like the initial segment of the following word, that is, a segment changes in anticipation of the articulation that is to follow. This type of assimilation is also referred to as **regressive** assimilation since the assimilation is in some sense "moving backwards" from the start of the new word to influence the end of the previous word. Alternatively, the initial segment of the second word may change to become more like the preceding context, a process referred to as **progressive** assimilation. Both regressive and progressive assimilation occur in English: the assimilation of [z] to [ʒ] when *is she* is pronounced as [ɪʒ ʃi] is an example of anticipatory (regressive) assimilation; the assimilation of [j] in *did **you*** to the [ʒ] component of the affricate in [dɪʤu] illustrates the process of progressive assimilation.

Assimilation may involve changes in voicing, nasalisation, or place of articulation. This thesis focuses on the last of these categories. In English, place assimilation most frequently involves alveolar segments (e.g. [t, d, n]) assimilating in the context of non-alveolars such as labials (e.g. [p, b, m]) and velars (e.g. [k, g, ŋ]). The direction of this type of assimilation in English is always anticipatory and is asymmetric: alveolars may change when preceding labials or velars, but the non-alveolars do not change preceding an alveolar context. So, for example, while the noun phrase *wooden basket* may be pronounced as [wʊdəm bɑskɪt], and the phrase *wooden casket* as [wʊdəŋ kɑskɪt], the phrase *ice-cream tub* would not be pronounced as [aɪs krin tʌb].

It should be noted at this point that assimilation is not necessarily an "all or nothing" process. Rather, assimilation may be either complete or partial (Barry,

1985; Kerswill, 1985). Indeed, the distinction between complete assimilation on the one hand, and partial assimilations involving evidence of residual articulation of the unassimilated segment on the other, has given rise to considerable debate in the phonological literature on how best to model the process of articulation in general. As Nolan (1992) observes, the kind of 'gradual' articulatory behaviour revealed by EPG data is hard to accommodate in a standard feature geometry approach to phonology (see Goldsmith, 1976; Clements, 1985, and others). Nolan acknowledges that autosegmental phonology provides a view of assimilation which supports the idea of articulatory economy:

> *"the autosegmental mechanism of deletion and reassociation seems more in tune with an intuitive conception of assimilation as a kind of programmed 'short-cut' in the phonetic plan to save the articulators the bother of making one part of a complex gesture." (Nolan, 1992, pages 262-263)*

But although escaping the constraints of a strictly segmental model – in which assimilation must be treated as a segmental **substitution** – the autosegmental approach "still portrays assimilation as a **discrete switch** from one subset of segment values to another" (*ibid.*, emphasis mine). It cannot account for the *range* of forms observed. Even when a more sophisticated autosegmental structure is adopted, whereby the place node dominates three separate *articulator nodes*, as proposed by Hayes (1992)[13], (accounting, in this way, for 'complex' double-articulated segments with more than one place feature) the problems of linear sequencing and the relative weakening of the residual alveolar articulation remain to be explained satisfactorily. Hayes' proposal does not explain why the first component is often only residually articulated (Holst and Nolan, 1995), for example.

Articulatory phonology, on the other hand (Browman and Goldstein, 1989, 1990), is readily able to account for a continuum of assimilated forms. Within this framework, linguistic structures are represented in terms of coordinated articulatory movements, called **gestures**, that are themselves organised into a **gestural score** that resembles an autosegmental representation. Different gestural types, such as *Velic gesture*, *Tongue-Tip gesture*, or *Lip gesture* are associated with particular articulators (in this case, the velum; the tongue tip, tongue body and jaw; and the lips and jaw, respectively), and each gestural type is represented on an independent tier of the gestural score. The theory assumes that continuous movement trajectories can be analysed into a set of discrete, concurrently active

[13] after Sagey (1986) and Ladefoged and Maddieson (1989)

underlying gestures. Much coarticulation and allophonic variation occurs as an automatic consequence of gestural organisation (see also Fowler, 1980). Browman and Goldstein maintain that all casual speech alternations result from two kinds of variation in gestural score: an increase in temporal **overlap**, and/or a decrease in the **magnitude** of a gesture (in both space and time). Importantly, they distinguish between overlap across and between articulatory tiers. In the former case, overlap gives rise to a **sliding** of events in time, whereas overlap on the same tier will result in a **blended** output trajectory (Browman and Goldstein, 1989, page 219).

However, recent analyses of [s]-[ʃ] assimilation by Holst and Nolan (Holst and Nolan, 1995; Nolan *et al.*, 1996) present data which is difficult to interpret within a pure Browman and Goldstein framework. The initially persuasive account offered by articulatory phonology fails to account convincingly for either the spectral or durational properties of what Holst and Nolan deem "type D assimilations", which involve a spectrally stable period of friction which is more [ʃ] than [s] like. Essentially these assimilations have the spectral characteristics of the segment in the assimilation-conditioning context (in this case, palato-alveolar) but the duration value of the unassimilated (alveolar) segment. Holst and Nolan conclude that two separate processes must be appealed to, in order to account fully for their data. In some instances, [s] undergoes articulatory blending with a following [ʃ] to yield a **contour** segment, a process which can be modelled most parsimoniously by gestural overlap. However, [s] may also become phonetically identical to a following [ʃ] *while retaining (at least part of) its original [s] duration.* The retention of durational values associated with the overlapped segment, [s], causes problems for an articulatory phonology account. The stability of the frication is indicative of complete gestural overlap, but such total overlap would also involve a change in duration to that of the overlapping segment, [ʃ].[14] Holst and Nolan conclude that assimilation processes of this type are better modelled as a phonological rule, whereby the [s] segment adopts the spectral characteristics of the [ʃ] while retaining its own duration value.

Cohn (1993) offers a similar treatment of nasalisation in English. She distinguishes between, on the one hand, the gradient quality of anticipatory vowel nasalisation, accounted for by a phonetic implementation rule within a target-interpolation model, and, on the other, the categorical nature of nasal deletion

[14]These results echo Ladefoged (1982)'s observation that assimilated segments may have different phonetic characteristics from their unassimilated forms, such as an absence of burst in assimilated oral stops, for example.

(and its interaction with coronal stop deletion and glottalisation), which is presented as an optional phonological rule.

In summary, assimilation is a process whereby a segment changes to become more like its preceding or following neighbour. The change may be complete, or only partial, and it is possible that different processes are responsible for the two alternative outcomes: partial assimilation is most succinctly accounted for in terms of gestural overlap within an articulatory phonology framework; the predictions made by current models of articulatory phonology fail, however, to capture some of the observed effects of complete assimilation, which is better explained in the traditional terms of phonological rule application.

2.5.2 Stop deletion

According to Heffner (1960):

> *"Fusion of morphemes in phrases frequently results in the* ***complete omission*** *of certain of the constituent elements. If the initial vowel of a morpheme is lost when it follows the final vowel of another, the grammarians speak of* ***aphaeresis****. When something is omitted from the interior of a morpheme, the phenomenon is called* ***syncope****, and when the final element of a morpheme is omitted, the term applied to the process is* ***apocope****."*
> *(Heffner, 1960, pages 178-179, emphasis mine)*

Although Heffner writes in terms of morphemes and their constituent 'elements', it is clear he is referring to the omission of what we would call phonological segments. Following Heffner's definitions, then, our concern here is with the process of deletion called **apocope**, specifically the deletion of word-final alveolar stops in English.

Cohen and Mercer (1975:301) characterise the stop deletion process in terms of the phonological rule in (2.6), where '#' denotes a *morpheme* boundary and '|' denotes a *word* boundary.

(2.6)

$$\left\{ \begin{matrix} t \\ d \end{matrix} \right\} \longrightarrow \emptyset \;/\; \text{obstruents}\ (\#) \_\_ \left\{ \begin{matrix} (\#) \\ | \end{matrix} \right\} \left\{ \begin{matrix} \text{obstruents} \\ \text{nasals} \\ \text{silence} \end{matrix} \right\}$$

It can be seen from (2.6) that they restrict the preceding cluster context to obstruents, and the following context to obstruents or nasals. Cohen and Mercer

also offer a phrase-final deletion option by including 'silence' within the set of possible following contexts. Given the restriction to preceding obstruents, Cohen and Mercer's rule does not account for instances of **homorganic** stop deletion, that is, the deletion of [t] or [d] when preceded by a nasal produced at the same place of articulation and followed by a consonant (with an optional intervening boundary). It also excludes the deletion of [t] and [d] preceding liquids and glides, such as in the phrases *vast lake* and *diamond ring.*

Oshika *et al.* (1975) observe that, at least for homorganic stop deletion it is possible that the rule might extend to [t] and [d] followed by a **vowel**, so long as the vowel is reduced, as in the word *twenty*, or the phrase *kind of*, illustrated in (2.7) and (2.8).

(2.7) *kind of* [kaind # əv] ⟶ [kainəv]

(2.8) *twenty* [twɛntɪ] ⟶ [twɛnɪ]

Stops are more likely to be deleted if they precede a nasal, sibilant, or [l] (Oshika *et al.*, 1975), so the homorganic stop deletion of [d] in the phrase *diamond mine* is more likely than the deletion of [d] in *gold ring.*

More recently McAllister *et al.* (1990:5) have specified the rule of stop deletion in English as follows:

> "**[t] or [d] occurring word-finally in a consonant cluster may be deleted when the following word begins with a consonant,** e.g. *vast meadow* may be pronounced [vɑs # mɛdoʊ] and *reclaimed fields* as [rikleɪm # fildz]."

McAllister *et al.* (1990)'s definition is less constraining than that of Cohen and Mercer (1975); it predicts the possible deletion of [d] in *gold mine* that is ruled out by (2.6) because the preceding cluster context [l] is not an obstruent, and the onset of the following word (*ring*) starts with neither an obstruent nor a nasal.

It is not clear that stops in any of these environments are in fact fully deleted, although this is implied by the notation of phonological rules such as (2.6) above. It is likely, that, rather, the stop undergoes processes of shortening, and where appropriate devoicing, and is left unreleased. Consequently, although the reduced stop itself may be difficult to perceive there may well be durational evidence in, for example, the preceding vowel, to cue the existence of an 'underlying' stop (see the discussion in Dalby, 1984). It is possible that close inspection of time-amplitude

waveforms and spectrograms might reveal the presence of short, unreleased stops where naive subjects, listening to the same data, report deletion.

2.5.3 Schwa syncope

According to Lindblom, vowel reduction is "a characteristic feature of languages with heavy stress, such as, for instance, English and Swedish" (1963:1773). Traditionally, vowel reduction has been equated with articulatory and acoustic **centralisation**, in other words, with movement towards the central or neutral vowel position generally associated with schwa (see, for example Tiffany, 1959; Shearme and Holmes, 1961).

Although all vowels in connected speech are subject to varying degrees of reduction, the role of **stress** in conditioning vowel quality has been widely documented in the literature (see Bates, 1995, for a detailed discussion). Unstressed vowels are significantly more reduced relative to stressed vowels, whether the term 'stress' is used to refer to lexical stress or sentential accent placement. In other words, unstressed vowels are less likely to reach **target** formant values, that is, they exhibit greater **undershoot** (Lindblom, 1963; Gay, 1978). Indeed, lack of spectral reduction is a cue for perceived stress (Rietveld and Koopmans-van Beinum, 1987). Unstressed vowels are also shorter and less intense than their stressed counterparts (Tuller *et al.*, 1982).

The duration of unstressed vowels depends on the position of the vowel within the word: word-final schwa is longer than word-initial schwa which in turn is longer than word-medial schwa (Umeda, 1975). Since schwa is characteristically short, one consequence of further shortening is the deletion of schwa altogether (Shockey, 1974), or, to use Heffner (1960)'s term, **schwa syncope**.

Dalby (1984) observes that schwa syncope has potentially significant repercussions:

> *"Of all the reductions that are characteristic of fast speech [schwa deletion] appears to be of central importance since it alters the syllable structure of words. Words that contain a schwa, front schwa, or syllabic sonorant in their underlying or careful speech forms are likely, in fast speech, to lose the nucleus of that syllable. If there are consonants that belong to the same syllable that are not deleted along with the nucleus, they must be resyllabified with the syllables that remain."*
> *(Dalby, 1984, page 9)*

Zwicky (1972) distinguishes the rule of **post-stress** syncope, which deletes the medial schwa in words such as *definite*, *mystery* and *general*, from **pre-stress** syncope, which deletes the nucleus of the initial syllable in words starting with a metrically Weak syllable, such as *below*, *banana* and *photography*. Zwicky argues – from his own intuitions of 'acceptability' – that there are three reasons for drawing the distinction:

1. There are words for which the post-stress rule is obligatory or has become lexicalised, while the pre-stress rule is never obligatory;

2. The post-stress rule applies according to the sonorancy of the following consonant, whilst the pre-stress rule applies everywhere in fast speech except when the application of the rule would result in 'unpronounceable' sequences of consonants;

3. Post-stress deletions are graded along the sonorance continuum and the tempo/style continuum, while pre-stress deletions are only acceptable in very fast speech.

Of course it is easy to disagree with intuitions, and some of the words which Zwicky claims involve an obligatory deletion clearly do not have the same character in all English dialects. The following words, for example, which for Zwicky would have only two syllables, in my idiolect, at least, would be trisyllabic in slow reading or citation form: *family*, *celery*[15], and *mystery*.

Dalby prefers to adopt a more objective approach by collecting and analysing a corpus of naturally occurring speech at varying tempos. In the first of two experiments Dalby transcribed and analysed conversational speech recorded from a televised news/interview programme. The second experiment analysed the output from three subjects who read a set of 183 test sentences at two tempos: slow (or 'normal reading speed'), and 'very fast' (subjects were instructed to read "as fast as they could say [the sentences] and still produce utterances they felt were possible and acceptable in a context in which they might be speaking rapidly" (Dalby, 1984:38)). In so doing, Dalby was able to observe both the linguistic facts and the frequency of occurrence of syllable deletion rules.

Dalby argues that the factors that condition the application of the syncope rule are more complex than Zwicky's claims suggest. Pre-stress deletion is clearly not

[15]even in fast speech I cannot apply the deletion rule to this word

restricted to fast speech: Dalby found that pre-stress and post-stress syncope occurred with about equal frequency in his corpus of conversational speech, that is, speech at 'normal' rather than exceptionally fast tempo. In general, word-medial post-stress syllables had the highest deletion rate, and word-final post-stress syllables the lowest, with word-initial (pre-stress) deletion rates falling between the two. This was true of both conversation and fast read speech, but not for the slow reading condition, where there was very little pre-stress deletion, and consequently no difference between pre-stress and word-final deletion rates.

In addition, analysis of adjacent segment effects did not support Zwicky's claim for a sonority effect of following consonant, although similar effects to the sonority prediction for preceding consonant context suggest that sonority may play a part, and that it is Zwicky's formulation of the effect that is incorrect. To the extent that sonority affects the likelihood of schwa being deleted, its influence is found to affect both pre- and post-stress syncope alike.

In the analysis of fast speech, consonant clusters adjacent to unstressed vowels were found to lower the percentage of deletion, relative to environments with no clusters. Furthermore, when a simple segmental parsing algorithm was run over the fast speech transcriptions, the results showed that unstressed vowels were about twice as likely to be deleted in environments where the remaining consonants could be resyllabified into onsets or codas that occur in careful speech forms than where they could not. Dalby concludes from this that the schwa deletion rule is a rule of syllable structure rather than one which refers to features of single segments. Since not *all* fast speech syllables were found to be well-formed, the syncope rule must be formulated in terms of a strong but not absolute tendency to delete schwa more frequently in 'parsable' environments.

Chapter 3

Lexical Representation and Process

3.1 Introduction

In Chapter 2 I introduced the problem of speech variability and outlined Lindblom's theory of hyper- and hypo-articulation which he offers as an explanation for the variation found in speech production. I highlighted two omissions in Lindblom's theory. Lindblom's failure to provide an explicit account of a listener's informational requirements will be addressed in the following chapter. Here, I focus on the issue of *distinctiveness*, the teleological component of the H & H theory that prevents speakers from reducing articulatory effort to its logical endpoint: silence. In particular, I consider the effects of lexical competition on a speaker's likelihood of hypo- or hyper-articulating. Lindblom's theory implies that speakers reduce articulatory effort while listeners are able to distinguish the input from other, similar sounding, words. I start by outlining the basic process underlying the successful recognition of words, and discuss the representational problem that variability in token production introduces. I describe briefly two theories that dominate the current literature on spoken word recognition: TRACE and the Cohort model, and then detail some of the empirical research which bears on the problem addressed by this thesis: the relation between variability in production and lexical competition.

Much of the work on spoken word recognition derives from earlier studies on the recognition of **written text**. As Lively *et al.* (1994) observe:

> *"Despite the unique nature of the speech signal and the issues it raises for speech perception, the bulk of research has been conducted using auditory analogues of tasks that were originally developed to study*

visual word recognition." (Lively et al., *1994, page 268)*

As a result, the inherent physical variability of the speech signal is sometimes overlooked. Because the type-written form of a word is unvarying, the problems introduced by variation in articulation are frequently ignored, as if it were assumed that listeners simply process well articulated citation forms. But as was observed in Section 2.4.4 variability in production may have significant consequences for how we model the process of auditory word recognition.

3.2 Some basic concepts

Spoken word recognition is not simply a matter of speech perception: of identifying or perceiving all the sounds of a language. Recognising the words which these sounds combine to form is not an inevitable result of identifying the constituent parts. We may perceive all the sounds and yet fail to recognise the word, as evidenced by the literature on **late** or **failed** recognition (see Grosjean, 1985; Bard *et al.*, 1988). Conversely we may recognise the word without hearing all of the sound components. This happens when a word is recognised **early** (Warren and Marslen-Wilson, 1987, 1988), or if some of the segments are **missing**, when a process of **phoneme restoration** results in the listener 'reinstating' the missing information (see Samuel, 1987, for example).

The successful recognition of a spoken word depends not just on identifying the component segments, but also on a number of factors outwith the acoustic details of articulation. In this section I introduce some basic terminology associated with the literature on word recognition, along with some of the key factors known to affect recognition rate and success: word frequency and the size of the competitor set.

3.2.1 The mental lexicon

Central to any theory of spoken word recognition is the concept of the **mental lexicon**. The mental lexicon is frequently likened to a dictionary which contains information about a word and its meaning. Precisely what information is or is not available via the lexicon is open to debate, but generally the mental lexicon is hypothesised to contain information about a word's phonology or sound structure, its orthography, some record of the syntactic roles which the word may take on, and a representation of meaning. The challenge for any model of spoken word

recognition is to characterise the way in which listeners use the incoming speech signal to access information stored in this mental lexicon.

3.2.2 Stages of lexical processing: access vs. recognition

Terminology in the spoken word recognition literature is frequently confusing and contradictory, with different researchers using the terms **lexical access** and **word recognition** to refer sometimes to the same, and at other times to different processes. In the discussion below, I will follow Frauenfelder and Tyler (1987), using the term 'lexical access' to refer to the period when information about lexical representations is made available, and 'word recognition' to refer to the outcome of a selection phase.

Most theories of word recognition assume that lexical processing starts with an **initial contact phase**: the listener takes the speech wave as input, and generates from this a representation which then makes contact with the internally stored form-based representations associated with each lexical entry. Clearly the nature of the input representation has important consequences for *which* lexical entries are initially contacted. The richer or more discriminative the information in the input representation, the smaller the number of lexical entries initially contacted.

When a contact representation matches some criterial part of a lexical entry the lexical entry changes state: it becomes **activated**. Different claims are made by different models about the relative status of activated words. In particular, models differ with respect to the quantity and type of information about the lexical candidates which is made available to the language processing system on activation.

Following the phase of initial contact is the **selection** phase. After a subset of the lexicon has been initially contacted and activated, accumulating sensory input continues to map onto this subset, until eventually one lexical entry is selected. The selection phase has been described variously as a process of *differentiation* (McClelland and Rumelhart, 1986), *reduction* (Marslen-Wilson, 1984), or *search* (Forster, 1976). The end-point of selection is when the listener has determined which lexical entry was actually heard: that is, the word is **recognised** (Frauenfelder and Tyler, 1987).

The term **lexical access** is used to refer to "the point at which the various properties of stored lexical representations – phonological, syntactic, semantic, pragmatic – become available" (Frauenfelder and Tyler, 1987, page 7). Theories

differ as to *when* this lexical information becomes available to the rest of the language processing system. For both TRACE (McClelland and Rumelhart, 1986) and the Cohort model (Marslen-Wilson and Tyler, 1980) lexical access precedes word recognition.

3.2.3 Defining lexical competitors

Almost all current models of word recognition involve the activation of multiple lexical candidates early in the recognition process (see Lively *et al.*, 1994). These multiple candidates have in common some part of the acoustic input. For example, the first syllable of the word *bandit* will activate a number of [ba] initial words, including *babble, backache, bad, badinage, ban, bank* and *battle.* When more of the target stimulus is heard, the activation levels of some of these alternative candidates will reduce, as the segmental composition begins to diverge from the input. Words that still match closely to [ban] include *ban, banish, banner, banister* and *banshee* as well as *bandit.* In TRACE, but not the Cohort model, other words will start to be activated, such as *an, animal, antelope* and *analyse.* All these activated word candidates are termed **competitors**: they compete with the target word for recognition.

Although TRACE and the Cohort model vary in how they define the competitor set for a particular lexical item, competition in both models is related to how **similar** the competitor is to the target word.

Words which share a large number of characteristics with the target are sometimes referred to as **neighbours**, especially with respect to the written form, where the term is usually reserved for words which differ by just one letter (Luce, 1986; Andrews, 1989; Luce *et al.*, 1990; Andrews, 1992). The spoken analogue refers to words which vary by a single segment. For example, visual neighbours of the word *band* include *sand, land, bend, bond, bald, bard, bank* and *bang.* Auditory neighbours, on the other hand, while including *sand, land, manned, fanned, bend, binned, boned* and *bond*, would exclude *bald, bard* (depending on the accent)[1] *bank* and *bang.*

The term **cohort** has been used (Marslen-Wilson and Welsh, 1978; Marslen-Wilson and Tyler, 1980) to refer to the **set** of competitors activated during the course of recognition, especially in relation to competitors that match at word onset i.e. to the first CV (see Section 3.4.2 below). In Section 3.5.3 I discuss the

[1] Although *band* and *bard* are not neighbours for non-rhotic accents such as R.P., for rhotic speakers who have no systemic difference between /a/ and /ɑ/ they will be.

empirical evidence that has been used to claim a special status for word onsets.

3.2.4 Word frequency

One of the main factors known to affect the recognition of a word is **word frequency**: how often the word form is encountered by the reader/listener (see, for example, Savin, 1963; Luce, 1986; Andrews, 1989, 1992)[2].

Frequency counts for words have traditionally been derived from analyses of large corpora of written texts (Thorndike and Lorge, 1944; Francis and Kucera, 1982), which it is assumed approximate the frequency distribution of spoken words[3]. Standard reference texts for word frequency record the frequency of occurrence of a given word per so many million words of text and number of text types, reflecting the prevalence of a word's use.

The effects of word frequency have been studied for several decades (see, for example Zipf, 1935). Savin (1963) demonstrated that listeners are able to recognise high frequency words at lower signal-to-noise ratios than low frequency words. In addition, incorrect responses to stimuli were always more frequent than the stimulus word. In other words, when subjects offered an alternative lexical item to the target, the word they offered was a higher frequency competitor.

Independent of the task used, high frequency words have been shown to be recognised faster and more accurately than low frequency words. Thus subjects recognise the word *bad* more quickly and accurately than they recognise the word *badinage*[4] while *banshee* will take longer to recognise than *banish*. High frequency words also require less acoustic information for successful recognition (Luce, 1984, cited in Lively *et al.*, 1994).

Frequency is an important consideration, therefore, in specifying the relative competition among items that are to be discriminated between (see Luce *et al.*, 1990).

[2]For a detailed discussion of issues relating to word frequency and its effects on various spoken word processing tasks see Kelly (1993) or Lively *et al.* (1994).

[3]Gradually, as more spoken language corpora become available on-line, it will be possible to derive reliable frequency measures for words as spoken, without this bias towards the written form.

[4]It should be noted that the difference in word length between *bad* and *badinage* will also contribute to the difference in reaction time.

3.2.5 Uniqueness versus Recognition Point

The exact point at which a given word is recognised depends on various factors, including its physical properties (duration, stimulus quality), its intrinsic properties (frequency) and the number and nature of other words in the lexicon to which it is similar, i.e. the number of competitors.

A number of researchers distinguish between a word's **Isolation** or **Uniqueness Point** (UP) and its **Recognition Point** (RP). The Uniqueness Point is defined as the point at which a word can be uniquely identified from its competitors, in other words, when the acoustic match diverges[5]. The UP may occur part way through the word, as in *abandon* which becomes unique at the onset of the first nasal i.e. [əban]; in some cases, however, a word may become unique only after its offset, such as *goal* which is contained within the longer word *gold.*

While the UP is usually treated as an unvarying property of a word-form in relation to its lexical neighbours, the Recognition Point varies according to factors such as word frequency and predictability. For example, rare words which form the onset of high frequency polysyllabic words (such as *hiss* in *history)* receive incorrect polysyllabic responses beyond the point at which the acoustic information diverges, that is, after the Uniqueness Point (Kelly, personal communication). Conversely, contextual support may result in a word being recognised before its UP. When presented with the sentence in (3.1) listeners do not depend upon hearing the final silence to recognise the word *goal* in preference to *gold.*

(3.1) *"Once Shearer passed the ball to him, there was no doubt that Gascoigne would score the winning* [gəʊl]"

3.3 Matching input to lexical form

Standard models of spoken word recognition assume some kind of mapping from the continuous speech waveform to discrete lexical units. I have already discussed the properties of the speech signal which make the recognition of words less than straight-forward to model (see Section 2.2.3). The variable, continuous and overlapping nature of spoken language has important consequences for the way in which speech can be processed. I raise here the issue of how such varying,

[5]This always presupposes that the listener knows where in the speech stream the word starts. While the segmentation issue can be ignored when investigating isolated word recognition it is clearly of relevance to the recognition of connected speech.

continuous input might be represented, and the process by which this input is matched to representations stored in the mental lexicon.

3.3.1 The issue of representation

Assuming that the recognition of spoken words involves a matching process between sensory input and an entry in the mental lexicon, the speech stream needs to be segmented and represented in a way which permits a correspondence between input and stored representation, that is, the two representations need to store the same kinds of information structure, or, in Frazier's terms, they need to be "couched in the same vocabulary" (Frazier, 1987, page 184). So, what information should be represented, and how? The nature of the representation will clearly have consequences for the matching process itself since it determines what information is being matched to what. The issue of the nature and form of representation is relevant to both sides of the matching process: i.e. both the input and target representation.

Any kind of representation necessarily requires an abstraction away from the original, with a consequent loss of detail (see Klatt, 1989, for more detailed discussion). The choice of representation level is therefore paramount. A phonemic representation, for example, abstracts away from acoustic-phonetic detail which may potentially be informative. Retaining too much detail, on the other hand, may serve to obscure an otherwise clear relationship, or else lead to computational expense in terms of storage and/or processing.

Psycholinguists working on spoken word recognition have tended, until recently, to work with a very traditional view of the lexical representation of words as listings of linear strings of phonemic labels.

As Frauenfelder and Lahiri (1989, pages 320-321) observe:

> *The psychologists' ignorance of basic phonological concepts is reflected by their view of input and lexical representations and by the way they have ignored problems posed by phonological processes. Until recently most models of lexical access have assumed extremely simple input representations consisting of linearly concatenated segments with no internal structure. These segments were often implicitly assumed to be phonemes; however, unlike the normal assumptions in the phonological literature, these phonemes were seen as wholes rather than as composed of features. Lexical representations were similarly conceived of as unstructured strings of segments. Lexical access was assumed to consist in a sequential mapping between these two representations,*

starting with the beginning of each.

Only quite recently has attention been directed to the issue of representation, with multi-tiered nonlinear representations of linguistic structure beginning to appear and be discussed in relation to the empirical evidence on lexical processing (Lahiri and Marslen-Wilson, 1991; Marslen-Wilson and Warren, 1994; Gaskell and Marslen-Wilson, 1996).

It is the *variable* nature of spoken language which lies at the core of the representation problem. Early work on spoken-word recognition tended to avoid or ignore the complexity of the sensory input, simplifying the problem by restricting research to the domain of isolated word recognition. Concentrating on the recognition of stimuli produced under laboratory conditions of careful and controlled articulation – that is, the recognition of citation forms – avoids some of the problems associated with the phonological reduction processes of connected speech. Thus, Frauenfelder and Lahiri (1989, page 320) observe that "the question of how listeners recognise different surface variants of the same word having undergone phonological processes has not received sufficient attention".

Phonological processes which involve the deletion, substitution, or addition of phonetic material alter the nature of the input, and consequently tend to obscure the relationship between input and lexical representation, complicating the mapping from one to the other (see the comments of Labov (1986)).

There are two basic alternatives to the problem of recognising multiple variable realisations as the same word form:

- all phonological variants of a word are listed in the recognition lexicon, i.e. all possible variation is "pre-compiled";
- only a *single* underlying representation is stored for any lexical item, which necessarily abstracts away from surface detail, but is compatible with all phonologically permissible variants.

The former solution is rather inelegant, computationally demanding, and dependent on the phonological theory being able to predict all the variation to be encountered. It also implies accurate listening abilities, with listeners perceiving all the available cues and then selecting the best fit from a list of similar options.

The latter approach necessarily requires an abstract level of representation that loses much of the surface detail. Listeners must use their phonological and

prosodic knowledge to recover the underlying lexical form from its diverse surface manifestations. The more perspicuous this recoding is, the less extreme are the demands made on accurate perceptual detail. Listeners contribute the rest of the information themselves, as Lindblom (1990a)'s theory proposes. Lindblom favours active listeners who use not just phonological and prosodic knowledge, but semantic and pragmatic information, too.

Clearly, given the variable nature of speech, the recognition lexicon must involve abstraction at some level, or else depend on some form of normalisation process prior to entry to the lexicon. Even if it were possible, it would be computationally crippling to have to store a representation for all and every possible realisation of any word form. The question is the degree of abstraction/normalisation required. How much detail must there be in the representation of a word-form in the mental lexicon for it to be recognised? The answer may best be sought in the research literature investigating the *process* of matching input to mental representation.

3.3.2 The notion of mismatch

Recently, research on the matching process has focused on the problems that arise through **mismatch**: when the contact input *fails* to match a lexical representation at some point.

There are two viewpoints from which mismatch can be investigated. Firstly, there may be a production error, a mispronunciation, which results in an articulation which is similar but not identical to any representation in the lexicon. A speaker says [ʃɪgəɹɛt] instead of *cigarette*, for example.

Usually, however, mismatch does not depend on a *mistake* in articulation: rather, mismatch can be viewed as an essential component in the definition of competitor sets. As observed in Section 3.2.3 above, a contact input may activate several lexical candidates that differ each from the other at some point in the representation. Mismatch relates to what happens to activated candidates once the discrepancy between input and some representations has been perceived. Mismatch results in a better fit for one candidate over another. Does mismatch then *block* the recognition of the mismatching word? If so, how does this process relate to the acceptance of a mismatch in a mispronounced word? Are some kinds of mismatch more acceptable than others, and if so, which?

Different models of spoken word recognition make different predictions about the effects of mismatch on recognition. Thus research on mismatch phenomena

may allow us to favour one model over the other on the basis of their ability to account for the empirical data. In the next section I describe two models of spoken word recognition which currently dominate the literature, outlining the basic principles underlying each model. I follow this with a presentation of some of the empirical evidence – offered in support of one or other model – which relates to the issues relevant to this thesis, *viz* the relation between acoustic mismatch, lexical competition and recognition.

3.4 Current models of spoken word recognition

3.4.1 TRACE

TRACE is an interactive model of spoken word recognition which emphasises the importance of top-down processing, that is, contextual effects (McClelland and Elman, 1986). It derives from McClelland and Rumelhart (1981)'s interactive activation model of letter and visual word identification, and, like any connectionist model, is implemented in the form of computer simulations. TRACE's interactive architecture conceptualises word recognition as a multi-level system in which activation at any level can affect activation in adjacent levels. This influence is **bi-directional**, allowing for both bottom-up and top-down influences on processing.

There are three levels in TRACE: the acoustic signal is processed into a set of input units, or nodes, which represent phonological **features**; these phonological feature units are connected to **phonemic** units, which in turn are connected to output units which represent **words**. When input units are activated the activation is propagated through the network, spreading from one level to another. As nodes at the featural level start to become activated, this activation will begin to feed through to the phonemic level and on to the lexical level. So, for example, when the nasalised vowel in the word *ran* is heard, the nasal feature node will be activated, which in turn will raise the level of activation of the nasal phonemes in English, /n, m, ŋ/. This will then activate lexical items such as *ram*, *ran* and *rang* (in addition to other words containing nasals such as *dangle*, *manage*, *mitten* *etc.*).

While links *between* levels are **excitatory**, the links *within* levels are **inhibitory**. All mutually incompatible units inhibit each other; thus the feature [+alveolar] will inhibit other feature place nodes; the phoneme unit /n/ will inhibit the

activation of other phonemes such as /t/ or /m/, and the word *ran* will inhibit its competitors like *run* and *rabbit*. Lateral inhibition has been characterised as a mechanism whereby "the rich get richer and the poor get poorer": highly active candidates damp down the activation of less potent competitors, exaggerating the difference between activation levels.

Returning to our earlier example, where the input had reached the nasalised vowel of the word *ran*, it can be seen that the activation levels of the competitors *rat* and *rag* will be lower than those for *ran*, *ram* and *rang* for two reasons: firstly, they will not have received the early activation resulting from the detection of nasality, and secondly, the more highly activated words *ran*, *ram* and *rang* will inhibit the excitation level of all lexical competitors.

The excitation of initially activated units changes according to the evidence of further input. If increased evidence enhances the probability that a hypothesis is correct, activation of that unit increases and the internal representation of the word hypothesis accrues further activation. Thus, as information about the place of articulation of the final consonant in the word *ran* is made available, there will be activation of the features which correspond to the alveolar [n] as opposed to labial or velar competitors, with consequent effects at the phonemic and lexical levels.

The amount of activation and inhibitory information from units at each level is determined by "the resting activation level of word units or [...] variation in the strength of phoneme-to-word connections" (McClelland and Elman, 1986, page 106). More frequent words are conceptualised as having higher resting levels, or stronger connections in the phoneme-to-word nodes than rare words.

As with all recognition models that generate multiple lexical hypotheses, there has to be a mechanism for declaring a winner. Within the TRACE framework, activation and inhibition continue until one word unit is consistently more highly activated than all its competitors. A winner is declared when one candidate commands a previously stipulated proportion (e.g. 0.9) of all activation (the "Luce decision rule" (Luce, 1959)). The greater the amount of overlap between competing hypotheses, the stronger the inhibition, but once one candidate begins to emerge the process of lateral inhibition should work in its favour, subduing its competitors. The primary effect of lateral inhibition, then, is one of **timing**: it tends to speed up the process of recognition or slow it down, according to the amount of match between competitors.

In a model like TRACE, goodness of fit takes into account the *overall* amount of

overlap between the input and lexical representation. The word *ran* will match the input string [ɹan] better than its competitors because it overlaps completely, whereas its competitors, such as *ram*, *rang*, *rat*, *tan*, *man*, and *run*, fail to match at some point during the word. Similarly, the input string [ʃɪgəɹɛt] will match the target word *cigarette* because the total overlap is greater for *cigarette* than any other lexical candidate, despite the production error at word onset.

However, TRACE will still show 'priority effects' owing to the intrinsic temporal dimension of spoken language. Word candidates which match the input at onset will have a 'head-start' over competitors which match only later in the word. For the input string [ʃɪgəɹɛt], therefore, word-initial matches such as *ship*, *shimmer* and *chivalry* will be more highly activated than *cigarette* at first, which, though activated by the lax vowel [ɪ], will be inhibited by the better matching *shi–* words. Only towards the end of the input will *cigarette* emerge as the best-fit candidate, the lateral inhibition it receives from the early matches delaying its emergence. So, while lateral inhibition boosts the recognition rate of words which match early in a word, subduing its competitors, it will slow the recognition of words which fail to match at onset and only later emerge as best fit candidates.

3.4.2 The Cohort model

The central idea underlying the Cohort model of spoken word recognition (Marslen-Wilson and Welsh, 1978; Marslen-Wilson and Tyler, 1980; Marslen-Wilson, 1984, 1987, 1989) is that as listeners hear the incoming speech signal, they set up a group – or **cohort** – of possible lexical items the word might be. This list of potential words is gradually whittled down until only one candidate remains. In Marslen-Wilson's own terms:

> *The process begins with the multiple access of word candidates as the first one or two segments of the word are heard. All the words in the listener's mental lexicon that share this onset sequence are assumed to be activated. This initial pool of active word candidates constitutes the* ***word-initial cohort****, which represents the primary decision space within which the subsequent process of selection will take place. The selection decision itself is based on a process of successive reduction of the active membership of the cohort competitors. As more of the word is heard, the accumulating input pattern will diverge from the form specifications of an increasingly high proportion of the cohort's membership. This process of reduction continues until there remains only one candidate that still matches the sensory input – in activation terms, until the level of activation of one recognition*

> *element is criterially discriminable from the level of activation of its competitors. At this point the form-based selection process is complete, and the word form that best matches the speech input can be identified.*
> *(Marslen-Wilson, 1989, page 7)*

In its original conception (Marslen-Wilson and Welsh, 1978; Marslen-Wilson and Tyler, 1980), the Cohort model required an exact match to CV onset. Thus, given the input [taɪm], the subset of activated hypotheses generated by the Cohort model would include *type*, *tide*, *tidy*, and *tycoon*; words such as *dime*, *rhyme*, *team* and *tame*, would be excluded since they fail to match the initial CV, despite the fact that these words share the same 'number' of correct features (phonemes) with *time* as words like *tide* and *type*. The difference is that for the *'dime'* set, the matched information is spread over the word as a whole, rather than concentrated at the onset.

The Cohort model's strict adherence to word onsets has been frequently criticised on the grounds that it assumes that the acoustic signal will provide unequivocal evidence for determining the word-initial segment from which to activate the cohort. The difficulties of segmenting continuous speech mean that a rigid specification of word-initial elements is simply untenable (see Section 3.5.3).

In theory the characteristics of the 'first pass' cohort are immune from contextual influence; the initial cohort is based on *sensory* information alone, with no effect of top-down information. Although this restriction on contextual influence has been criticised, evidence from gating and priming studies (see Sections 3.5.1 and 3.5.2 below) appears to support the Cohort model's predictions (Tyler and Wessels, 1983; Tyler, 1984). For example, Tyler and Wessels (1983) show that when presented with target words in context, listeners' guesses at early gates (i.e. before the first 200 ms of a word) while mostly compatible with the sensory input were often contextually inappropriate. However, McAllister (1988) presents contradictory evidence which reveals that initial guesses do depend on the context.

The Cohort model has no direct mechanism for dealing with lexical competition. Because lateral inhibition does not apply, the presence of a competitor does not directly affect the activation level of a word candidate (Marslen-Wilson *et al.*, 1996). Rather, activation is the result of the goodness of fit between the input and the lexical representation, that is, the amount of information overlap, or degree of match. However, although there is no lateral inhibition, there appears to be inhibition at the phonetic level. Bottom-up inhibition arises as a consequence of mismatching input having a direct negative effect on the activation level of the

target representation. Thus hearing [pleɪn] as input will result in the word *plate* being inhibited – as a result of the nasality mismatch – and consequently ruled out as an interpretation. This effect is entirely independent of the simultaneous activation of the word *plane*, which has no direct effect on what happens to the activation level of *plate*.

Recent versions of the Cohort model incorporate lexical neighbourhood effects by proposing that competition between candidates will result in a **delay** in recognising the target word: "Recognition depends on differentiating the activation of a candidate from the activation levels of all its competitors, and the more strongly its competitors are activated, the longer the delay until the correct word is securely identified" (Marslen-Wilson *et al.*, 1996, pages 5-6). Since strongly activated hypotheses stay highly activated for longer, delays may occur before the activation levels of really strong competitors decay sufficiently to achieve a criterial difference in activation for the target.

3.5 Assessing the empirical evidence

In the following section I outline some of the research that has been used in support of one model over the other. I concentrate on empirical work which has a bearing on the investigations presented in this thesis, namely

- dealing with phonological variation, especially assimilation;
- the issue of failing to match at word onset.

The empirical evidence comes primarily from two experimental paradigms: **gating** (Grosjean, 1980; Tyler and Wessels, 1983), and intra- and cross-modal **priming** using both repetition (Monsell, 1985) and associative primes (Moss and Marslen-Wilson, 1993). I shall describe briefly the methodology behind these two techniques before detailing some of the empirical results they have generated.

3.5.1 The gating paradigm

In a gating task listeners are presented with increasingly large fragments – or **gates** – of stimuli. At each presentation the listener responds by stating what they believe the stimulus to be, in some cases adding a confidence rating. Gating has been used primarily for presenting isolated words, starting, for example, with the first 50 ms from onset. The second stimulus contains all the speech that was

heard before, plus the next 50 ms, the following gate contains the previous 100 ms plus another 50 *etc.* until the final gate when the whole word is presented. Gating 'through' a word like this allows the experimenter to assess how the listener is interpreting the sensory information presented up to the point at which the current gate ends. It is possible therefore to show what types of word candidates listeners generate in the course of on-line processing of speech inputs, and to explore the timing with which candidates emerge from their competitors.

Since the term 'gating' simply refers to the incremental presentation procedure there is no requirement that the stimuli are necessarily word fragments nor that they are presented in isolation. It is possible to use the technique to present words, one at a time, from the start to the end of a phrase or utterance, or, for example, to present a sentential context up until the onset of the word of interest and then begin gating through the word at that point. Similarly the size of incremental chunk can be varied, though for word gating it is usual to take gates that are greater than a pitch period but smaller than the length of a segment, that is, roughly 50 ms or so.

3.5.2 The priming paradigm

The principle underlying priming studies is that presenting a related stimulus – or **prime** – at some point prior to a target stimulus to which the subject must respond (the probe) leads to a **facilitation** of the target response.

The response task is usually to make a lexical decision ("Is this a word, yes or no?"), and the dependent variables are consequently Reaction Time (RT) and number of errors made. A significant decrease in RT from a control base-line indicates a priming effect, or facilitation of the target word.

The prime is necessarily related to the target in some way; it may be the identical (or source) word, as in **repetition priming**, or it may be semantically or associatively related to the target, called **semantic** or **associative priming**. An example of the latter would be the use of the prime *honey* to facilitate responses to the target BEE. Semantic relations tend to involve hyponymy (e.g. *rose—flower*), or semantic properties (e.g. *sour—lemon*), while associative relations reflect frequent co-occurrence (e.g. *cup—saucer*). The strength of such relationships is usually pre-tested in property generation, free association, and/or multiple cloze tests (see, for example, Moss and Marslen-Wilson, 1993).

Presentation of prime and target may be either intra- or cross-modal. In intra-

modal priming experiments, both prime and target are presented in a single modality, either auditory or visual. In cross-modal priming studies, primes are usually presented in the auditory modality with responses required to visually presented targets. It is a convention in describing material for these types of experiments that visual probes are written in upper case, while auditory stimuli are written in italics. Thus BEE would be a visual target, while *bee* would refer to its auditory prime (in this case, in a cross-modal repetition priming condition).

3.5.3 Location of mismatch: how important are word onsets?

One of the principle differences between the Cohort model and TRACE is the emphasis placed by the Cohort model on word onsets. The Cohort model advocates a 'decision space' determined by the beginnings of words. In the strictest form of the model, failure to match at onset will result in a word not being recognised at all. This contrasts with the TRACE model of word recognition where the *overall* goodness of fit between activated candidates and the input representation may rescue an early mismatch.

In a paper designed to address the issue of whether *overall amount* of match mattered more than the *location* of match, Marslen-Wilson and Zwitserlood (1989) examined the extent to which partially matching input produced activation.

Earlier research (described in Marslen-Wilson & Zwitserlood, 1989) revealed that word-initial partial matches were successful at activating lexical representations sufficiently to produce priming effects in a cross-modal study. Subjects were presented with complete spoken words such as *kapitein* (*captain* in Dutch), which share a large amount of initial overlap with a second word such as *kapitaal* (meaning *capital*). At the same time as subjects heard these words, they were presented with a visual probe to which they had to make a lexical decision response. The timing between auditory and visual presentation was varied, so that the visual probe was either presented at the offset of the auditory prime, or else while the word was still being heard, at a point when "the input was still compatible with both possible words, for example, before the release of the [t] in *kapitein* or *kapitaal*" (Marslen-Wilson and Zwitserlood, 1989, page 577). Visual probes were semantically related to one or other of the two *kapit*-initial words: BOOT (*ship*) for *kapitein* or GELD (*money*) for *kapitaal*. Analysis of RT demonstrated that when the probes were presented in the middle of the spoken word then both probes were facilitated, independently of which word the auditory stimulus turned out

to be. In contrast, when the probes were presented at the end of the word, only the probe related to the word heard was facilitated. In other words, by the end of the word *kapitein* there is priming for BOOT but not for GELD.

A subsequent set of experiments (Zwitserlood, 1989) used *fragments* of spoken words presented this time in sentential contexts. A mean facilitation effect was found for probes related to the word from which the fragment was taken, for example to the probe BOOT when the fragment [kapi:] was taken from *kapitein.* A facilitation effect was also found for probes related to the fragment's close competitor, in this example, to the probe GELD (which is related to *kapitaal*). Thus word-initial fragments appear to match full lexical representations and activate them sufficiently to create a priming effect.

The question addressed by Marslen-Wilson and Zwitserlood (1989) was whether partial matches could be non word-initial and still facilitate targets. Using cross modal associative priming in Dutch, they compared the priming effects of an original source word with both real and non-word **rhyme primes**. For example, as primes to the visual target BIJ (English = BEE) they compared the source word *honing* (*honey*), the rhyming real word *woning* (*dwelling*) and the non-word *foning*, in addition to a control word *pakket* (*parcel*), and a control non-word *dakket.* The equivalent paradigm in English might use *honey*-BEE, *money*-BEE and *shunny*-BEE with suitable controls.

Marslen-Wilson and Zwitserlood found that in general rhymes such as *woning* and *foning* failed to prime the target BIJ. In fact, rhyme primes produced significant facilitation only when the competitor environment of the source word was particularly sparse. In addition, no effect of lexical status was found: real and non-words were equally ineffective as partial primes.

In *post hoc* analyses, Marslen-Wilson and Zwitserlood – adopting a simple segmental analysis of the input string – grouped items according to the amount of segmental overlap between source and prime. The groups ranged from three overlapping segments (as in *herrie/merrie/lerrie*) to six or more segments overlapping (as in *handelen/wandelen/janderlen*). No effect of overlap length was found: the advantage the source word held over the rhyme primes remained constant as the amount of matched input varied.

Marslen-Wilson and Zwitserlood argue from these results that word onsets clearly *do* have special status in spoken word recognition: even in the maximal overlap conditions the rhyme prime never catches up with the original word. Such results, they claim, are inconsistent with a simple global notion of the nondirectional

mapping of stimulus information onto lexical representations[6]:

> *"If the simple amount of matching input is the critical variable in determining amount of activation, then a rhyming stimulus like* woning *should facilitate responses to* BIJ *at least as much as a stimulus like [kapi:] should facilitate responses to* GELD *and* BOOT*" (Marslen-Wilson and Zwitserlood, 1989, page 578).*

However, as Tabossi (1993) points out, Marslen-Wilson and Zwitserlood are not comparing like with like. Whereas the initial fragment [kapi:] is shared between two different words, that is, the offset of the fragment is prior to the Uniqueness Point of any one word, the rhyme primes like *woning* are complete, well-formed words in Dutch, which are semantically unrelated to the target word. A better comparison would be the earlier data discussed in Marslen-Wilson and Zwitserlood where the target word is presented *after* the prime's Uniqueness Point; in this case, as in the bulk of the semantic priming literature, the primes no longer facilitate the targets associated with their competitors, that is, *kapitein* does not prime GELD.

There are at least two possible reasons for Marslen-Wilson and Zwitserlood's failure to find rhyme priming. The first is that it may not be the word-initial mismatch *per se* that blocks facilitation, but rather the better match to an alternative candidate: perhaps mismatch is problematic only when it leads to the increased activation of a competitor. Models like TRACE can account for *woning* failing to prime BIJ on the basis that the acoustic input matches the real word *woning* better than the similar but different *honing*. While this may account for the lack of priming for real word rhyme primes, it is not so clear why the non-word primes like *foning* are equally ineffective at facilitating the targets, unless, perhaps, they happen to match better to other real words (like *woning*) than to the target (*honing*).

Alternatively the results can be accounted for by emphasising the importance of information *mis*match. A crucial difference between the [kapi:] fragments and the rhyme primes is that while both cases involve considerable featural overlap between competing candidates, in the former case there is *no* mismatching information, while for the rhyme primes there is conflicting information about the identity of the first segment. Perhaps it is this conflicting information that is significant, independent of whether the mismatch occurs word-initially or later in

[6]This is the interpretation they place on models like TRACE, forgetting, clearly, that TRACE, too, incorporates a directional mapping with matches at onset receiving a 'head-start' over later matches.

the word (such as after the [t] of *kapitein*).

Furthermore, while rhyme primes are always less effective than the source at facilitating the target words, word-initial mismatch does not provide an *absolute block* to entry into the decision space: when the original word has only one rhyme prime as competitor, then the partial prime can lead to facilitation. At the very least this result argues against the strongest version of the Cohort model, which predicts that all rhyme primes (which by definition fail to match at word onset and cannot therefore be word-initial cohort members) will fail to facilitate the target, regardless of competitor strength.

Nooteboom (1981)'s investigation of lexical retrieval from word fragments confirms this view. Nooteboom compared the recognition rates of fragments of words taken from either the beginning or the end of a word. In all cases the fragments included just enough information to make them uniquely identifiable. For example, if the Dutch word *surrogaat* (/sœro:χ'a:t/) is split in the middle of the /o:/, the initial string /sœro:/ could be the start of only one word in Dutch, *surrogaat*, while *surrogaat* is also the only word that ends in /o:χa:t/. Nooteboom found that while word-initial fragments are more easily recognised than fragments taken from word endings, the probability of correct retrieval of word-end fragments was still 0.61, suggesting that subjects are sometimes able to recognise words without access to the initial CV onset.

3.5.4 Competitor effects on rhyme priming

Marslen-Wilson, Moss, and van Halen (1996) extended the work by Marslen-Wilson and Zwitserlood (1989) to explore the properties of the listener's lexical decision space: they asked whether phonological distance affects the acceptability of the prime as a token of the source word, and whether this is modulated by the presence or absence of close competitors in the lexical space. In an intra-modal auditory priming experiment, they varied the distance between the source and non-word rhyme prime (close, distant) in three competitor conditions: rhyme competitors were Close, Distant or Absent *vis-à-vis* the non-word rhyme prime.

Although they did find some rhyme priming, rhymes primed much less effectively than the source words themselves. It seems that even a single-feature deviation in the first segment of a word results in the system treating the input as perceptually distinct. Nevertheless, there was still significantly more facilitation from the non-word rhyme primes than the control base-line, providing stronger support for

rhyme priming than earlier studies (cf. Marslen-Wilson & Zwitserlood, 1989).

The effects of source distance and rhyme competitor were much weaker, with a marginal priming advantage for close over distant primes, and a suggestion of a possible competitor effect: priming effects looked strongest where the rhyme competitor is absent, and weakest where the rhyme prime is phonologically close to the competitor and distant from the source word. This effect, though significant in *post hoc* tests, lacked broader statistical support.

In a second experiment – this time using cross-modal priming in English – listeners were presented with rhyme primes that began with a perceptually ambiguous segment, created by manipulating the Voice Onset Time (VOT) of the word-initial stop consonant. In the Word Competitor group, the source word (e.g. *plank* with its associate WOOD) has a real word competitor (*blank*), and the ambiguous prime (*b/plank*) is therefore equally close to the two. In the No Competitor group, the source word (e.g. *task*, associate JOB) has no real word rhyme prime that differs only in voicing, so the rhyme competitor is a non-word (*dask*). In this case the ambiguous word is close only to one potential lexical target, the source word (*task*).

Reaction Time responses showed a significant prime type effect, with a significant interaction with competitor type. For the No Competitor items, the ambiguous prime behaved like the source word, producing significantly faster RTs to the probe than either the control base-line or the (non-word) rhyme prime (i.e. *d/task* = task). However, for the Word Competitor items, only the source word produced significant facilitation to the probe; the ambiguous prime behaved like the real word rhyme prime, which did not differ from the base-line control (i.e. *b/plank* = blank).

This outcome reinforces the view that the strongest form of the Cohort model is incorrect with respect to the mapping of sensory inputs onto lexical representations. Clearly, this mapping is not always blocked by word-initial mismatch, especially when the mismatch creates a non-word rather than another real word. Models such as TRACE, which allow for the mapping of [ʃɪgəɹɛt] onto *cigarette* are better able to account for such results than can the traditional Cohort model.

Marslen-Wilson *et al.* suggest that lexical context may only affect the interpretation of *ambiguous* stimuli and not the clear cases, such as the endpoints of place or voicing continua. But then one might ask how many such *clear* cases do listeners encounter in their day-to-day processing of real speech? Surely a model of spoken word recognition needs to account not just for the unusually clear cases,

but for the mucky tokens, too. Recently, attention has been turned to these more problematic tokens in an attempt to understand how listeners cope with words that have undergone phonological changes such as place assimilation. This, in turn, has refocused the debate on the issue of representation.

3.5.5 Matching underspecified representations

In an attempt to provide a linguistically adequate account of lexical representation, Lahiri and Marslen-Wilson (1991) present the results of a gating study which advocate an abstract underspecified representation of lexical form. They argue against an intervening segmental level, proposing, rather, that input to the lexical level is featural.

Underspecification (Kiparsky, 1982; Archangeli, 1988; Archangeli and Pulleyblank, 1994), although a development that has arisen from autosegmental approaches to phonology, can be seen to descend from SPE (Chomsky and Halle, 1968) where preference was expressed for a grammar in which only idiosyncratic properties are lexically listed and predictable properties are derived. According to underspecification theory in its radical form (Archangeli, 1988), every linguistic item has a single unique Underlying Representation (UR) which is **minimally specified** in its phonetic description. This UR is based on a hierarchical non-linear structure, with multi-dimensional information represented at various levels – or tiers – that are linked (cf. Goldsmith, 1976; Clements and Keyser, 1983; Clements, 1985).

Predictable features are derived by rule and therefore not specified in the UR. Predictability is defined by two principles: **redundancy** and **underspecification**. While redundancy determines *which* features are specified, underspecification determines the *value* of a feature to be represented. (See Steriade, 1987, for a discussion of redundant versus distinctive feature values.) Underspecification necessarily incorporates the notion of default value, or **markedness**, so that instead of representing values as binary distinctions [+/− feature] there is a third option, which is to leave the value as 'unspecified'. The only specifications in the UR are those for features which are a) distinctive and b) have the marked i.e. non-default value. Default values are left unspecified, with the feature value being filled in by the appropriate redundancy rule. It is this underspecification that has significance for lexical access, Lahiri and Marslen-Wilson (1991) claim.

The process of assimilation is offered as evidence for the principle of underspecifi-

cation: "if a class of sounds is not specified for a given feature, a feature-filling rule can spread a feature from a neighbouring segment" (Lahiri and Marslen-Wilson, 1991, page 253). Asymmetries in assimilatory processes, such as the tendency for coronals to assimilate in place to the following consonant while velar and labial segments remain unchanged, can be readily explained if it is assumed that coronals are not specified for the feature class [place] and can therefore 'borrow' the place feature from the neighbouring consonant.

Hypothesising that the lexical representations deployed in speech recognition also contain only distinctive and marked information, Lahiri and Marslen-Wilson argue that, if a feature is unspecified, then variation in the phonetic realisation of that feature will not affect the goodness of fit between the sensory input and the lexical representation. They tested this hypothesis by exploring listeners' interpretation of phonetic cues over time via a gating task.

The oral/nasal contrast for vowels was selected as the feature to study for three reasons. Firstly, the representation of nasal vowels as [+nasal] (i.e. marked underlyingly) and oral vowels as unspecified for nasality is uncontroversial. Secondly, there are two distinct sources for the surface feature 'nasalised vowel': vowels may be underlyingly nasal, or they may be nasalised as a result of regressive feature assimilation from a following nasal consonant. Lastly, a cross-linguistic comparison can be made between English and Bengali, since both languages have a rule of regressive nasal assimilation, but only Bengali has a systemic distinction between nasal and oral vowels.

The materials consisted of three sets of stimuli:

- 21 Bengali triplets of CVC, CVN and C$\tilde{\text{V}}$C words, such as [kɑp], [kɑm] and [kɑ̃p], using the five vowels [ɑ, o, æ, ɔ, e] and their nasal counterparts;
- 20 English CVC/CVN doublets, such as [kʌp] and [kʌm], matched to the Bengali words in phonetic structure and frequency (where possible);
- 20 Bengali CVC/CVN doublets, involving a 'gap' in the lexicon where there would otherwise be a C$\tilde{\text{V}}$C word, such as [lop] and [lom], since there are no [lõ] words.

Predictions concern the interpretation of nasalisation in the vowel prior to hearing the conditioning consonant that follows. Under a *surface* representation, a [+nasal] vowel will match onto both CVN and C$\tilde{\text{V}}$C words but will mismatch CVC words, while an oral vowel will match only an oral context (i.e. CVC words). If,

however, nasality is *underspecified*, then not being specified for the feature [nasal] means a stimulus can potentially match to the larger set of words which includes both nasal and non-nasal vowels. Thus, on hearing CṼ listeners may respond with CVC, since there is no *mis*match between the sensory input (which is nasal) and the underlying representation of the vowel which is unspecified for nasality. The only occasion mismatch will arise is for a CṼC response to a CV (non-nasal vowel) stimulus.

For all stimulus sets, responses to the *oral* vowel (CV stimulus) resulted in roughly 80% CVC judgements with the remaining responses CVN. For the Bengali triplets, for which CṼC was also an available real-word response, fewer than 1% of responses involved a nasal vowel (Ṽ). That is, a CVC stimulus rules out a CṼC response but accepts CVN responses. This result presents problems for a surface representation account, which would predict only CVC responses. CVN responses under this view are simply a mistake. Abstract underspecification, on the other hand, predicts both CVC and CVN responses since in either case there is no mismatch between the non-nasal vowel heard and the UR for the vowel which is simply unspecified for nasality. The ratio of CVC to CVN responses (roughly 4:1) is a reflection of the distributional properties of the language as a whole: because there are fewer nasal-final words in the language there are fewer nasal-final responses overall. This holds for both the Bengali and English response patterns.

Responses to the *nasalised* vowels varied according to which language was heard. For Bengali speakers there was a strong tendency to respond with CṼC words to both CṼC and CVN stimulus conditions. The presence of nasalisation was clearly treated as CṼ and not as CVN, that is, it was responded to as underlyingly nasal. Indeed, for the Bengali triplet data there were fewer CVN responses to the CVN and CṼC stimuli (7.9% and 5.2% respectively) than to the CVC (i.e. oral vowel) stimuli (13.4%). The CVN responses to CVN stimuli remain below 5% until the final gate before the vowel offset. With the first gate into the nasal consonant there is an immediate switch over to CVN responses as the nasal context becomes apparent and enables the listener to disambiguate the source of the nasality. Responses to the Bengali doublets demonstrate a reluctance to produce CVN words to CVN stimuli (15.6%) despite the fact that no CṼC response is available. Instead, the nasalised vowel prompted responses with CṼ structures, such as phonologically closely related words, or sometimes 'nonsense' words, which were potential real words in other dialects of Bengali, or other languages (like Hindi) which would have been known to the listeners.

The reluctance to produce CVN responses to nasalised vowel stimuli is inexplicable on the surface representation account: if a CVN vowel is represented as [+nasal] then it should match a nasal vowel in the sensory input. There is no such difficulty in accounting for the results via abstract underspecification. Under this view, the nasalised vowel input will match completely with the [+nasal] specification for the CṼC vowel, that is, CṼC responses will provide the *best* match. However, there will be no mismatch for CVN or CVC words (since these have no specification for nasality) so some of these would be expected as responses in addition to the CṼC words. Where there are no better fitting CṼC competitors available in the language, as in the Bengali doublet cases, CVN stimuli match both CVC and CVN responses in a similar way to the CVC stimulus situation.

English speakers, on the other hand, have no underlyingly nasal vowels to provide a 'better' match to the nasalised vowel input than an underlyingly oral vowel (CVC or CVN). Consequently, as soon as listeners start to detect nasalisation in the stimulus input, they can respond with CVN words, since vowel nasalisation will be interpreted as evidence about the properties of the following segment (cf. Warren & Marslen-Wilson, 1987). The results for English responses to CVN stimuli cannot of themselves distinguish between the two alternative representation hypotheses. On the surface representation hypothesis, CVN words are represented with nasalised vowels, and can be discriminated from CVC words as soon as vowel nasalisation can be detected, since there are no CṼC words to lead to ambiguity. Thus the outcome is the same whether the vowel in a CVN word is represented as [+nasal] or unspecified for nasality.

Lahiri and Marslen-Wilson (1991, page 291) argue from their results that "listeners do not have available to them, as they process the speech input, a representation of the surface phonetic properties of a given word-form." Rather, their performance is determined by the abstract, underspecified representation in the recognition lexicon. This representation must be radically underspecified in order to account for the Bengali data (cf. Archangeli, 1988, and Steriade, 1987). Contrastive underspecification would require the vowels in CVC and CVN words in Bengali to be specified as [−nasal] since the feature [nasal] is used to distinguish segments like [o] and [õ]; under such conditions CṼC stimuli would then mismatch CVC and CVN responses, contrary to the results of Lahiri and Marslen-Wilson[7].

The simplified representation advocated here by Lahiri and Marslen-Wilson, stripped of predictable and default information, offers simpler and more direct proce-

[7]and of Ingram & Mylne (1994), reported in Marslen-Wilson, Nix & Gaskell (1995), who replicated these results in a gating study of French.

dures, they claim, for understanding how to deal with variation. Phonologically variant surface forms can access the appropriate lexical entry directly, without requiring any intervening process of inference or canonicalisation. In the following section I show how Marslen-Wilson and others have used gating and cross modal priming studies to explore the effects of variation on lexical decision, and ask to what extent underspecification in the lexical representation can account for their results.

3.5.6 Mismatch as phonological variation

The primary processing problem posed by phonological variation is to prevent variant forms from creating mismatch and, as a consequence, failing to be recognised (Marslen-Wilson *et al.*, 1995). Can this problem be solved by assuming an abstract lexical representation in which only distinctive, non-redundant, marked information is specified?

One of the requirements of underspecification is that information is recoverable, that is, the underlying representation must be accessible from the surface form. For example, a vowel that is unspecified for nasality underlyingly can be nasalised in the surface form, but a vowel that is underlyingly nasal could not appear as phonetically oral, since there is no rule in the UG that can be applied to get back to the nasal form. Thus only unspecified features are liable to undergo feature-changing variation. This prediction appears to be borne out by research by Shillcock *et al.* (1996) who observe that phonological feature changes occur in the direction of a less frequent segment, following a 'Move-to-Markedness' principle. (See also the discussions on the status of coronals in Paradis & Prunet, 1991.)

Assuming phonological processes that result in feature change apply only to underspecified segments, then because these features are unspecified in their lexical representation, the variation that arises should not cause mismatch (as there is no specified value from which the speech input can deviate). While this appears to hold for the nasalised vowel data of Lahiri and Marslen-Wilson (1991), it is not clear that feature-changing variation which arises from processes applying across word boundaries will have the same effect. If, for example, the word *rat* assimilates in place to [ɹak] preceding the word *cage*, then the lack of mismatch to the velar place information would result in both *rack* and *rat* being activated, resulting in perceptual ambiguity. Of course in most circumstances such potential ambiguity will be resolved by various phonological, lexical and sentential

constraints. However, using pairs of stimuli for which such disambiguating constraints were not available, Marslen-Wilson, Nix and Gaskell (1995) attempted to test the relative ambiguity of word-final coronals and non-coronals.

In the first of two experiments, sentence pairs like those in (3.2a) and (3.2b) were presented to listeners who heard the entire left context and then incrementally larger portions of the critical word ([leɪt]/[leɪk]).

(3.2) a. *They thought the [leɪk] cruise was rather boring.*
(Assimilated)

b. *They thought the [leɪt] cruise was rather boring.*
(Unassimilated)

A forced-choice pre-test (where listeners were asked to choose between two possible readings (*late/lake*) for the [leɪ] fragment they were hearing) showed that the critical tokens did indeed contrast phonetically (86% of non-coronals were judged as non-coronal, while 93% of coronals were judged as coronal); these values were used as a base-line measure against which the gating responses were compared.

The gating task produced a significant drop in non-coronal (i.e. velar and labial) judgements for the assimilated tokens, which fell to chance level (52%). When the final segment was coronal there was a small but non-significant drop (to 80%) in coronal responses. In other words, when trying to recognise the word in context (rather than selecting from two forced-choice alternatives) subjects identified [leɪt] as *late*, but had problems recognising [leɪk] as *lake*, responding just as frequently with the coronal alternative *late*.

The difficulty listeners encountered in responding to the assimilated tokens in the gating task is interpreted by Marslen-Wilson *et al.* (1995) as suggesting that a phonetically unambiguous velar or labial gesture "can be re-interpreted as lexically coronal" (1995:296). Is this because the lexically coronal segment is underspecified for place in its abstract lexical representation? To test this hypothesis, a second experiment, this time using repetition priming, was conducted on isolated words, comparing facilitation of various primes to target words ending in coronal and non-coronal segments. Three types of auditory prime were contrasted: for a coronal target such as HEAT, primes were either identical (*heat*, [hit]), real words that mismatched at offset (*heap*, [hip]), non-words that mismatched at offset (*heak*, [hik]), or unrelated controls (*drop*). These were compared with primes to non-coronal targets like FAKE, which were identical (*fake*, [feɪk]), word mismatches (*fate*, [feɪt]), non-word mismatches (*fape*, [feɪp]) or control base-line words (*shed*).

No priming effect was found for the word and non-word mismatch conditions for either coronal final or non-coronal final words. The target *heat* was not facilitated by *heap* or *heak*, nor was FAKE facilitated by *fate* or *fape*. Given the possibility that the stop-release might have provided cues to the word having been spoken in isolation[8], the experiment was rerun with the consonantal release *completely* deleted. Again, there was significant identity priming and very little priming for the mismatch conditions. As Marslen-Wilson *et al.* (1995:300) themselves observe, "these results provide little comfort to a strong representational hypothesis": the strong version predicts that labial or velar place information will not mismatch a coronal target because it is unspecified for place underlyingly, yet even in cases of restricted place information (when the stop release is truncated) priming of the coronal word appears to be blocked.[9]

Perhaps an alternative explanation ought to be sought for the increase in coronal responses to assimilated tokens like *lake* presented in sentences such as (3.2a). A check on word frequency via the CELEX database[10] reveals that the coronal-final word *late* is the **most frequent member** of the [leɪ] cohort (with a frequency of 11,722), while *lake* is the sixth most frequent member (frequency = 889). In a gating task which requires subjects to recognise the incoming increments with little conditioning context to constrain the lexical search (*"They thought the"* could as likely be followed by *late, lady* or *Labour*) responses will clearly be biased towards the most frequent candidate. Since Marslen-Wilson *et al.* (1995) do not offer their dataset as an appendix it is not possible to check whether this frequency bias is reflected in the data as a whole, or is an artefact of the one example they present. It is quite plausible, however, that the results of the gating study could be accounted for in terms of simple word frequency, rather than by reference to underspecified representations.

If word frequency is the cause of the *late* responses to [leɪk] in the gating experiment, then how is the different response profile for the forced-choice pre-test

[8]Earlier studies (Marslen-Wilson, 1993) have shown that non-words derived from real words by changing place of articulation from coronal to velar (or labial) fail to prime targets that are semantically related to the source word. For example, the sequence [flik] derived from *fleet* ([flit]) did not prime lexical decision responses to SHIP. Results from gating (Marslen-Wilson *et al.*, 1995) suggest that one factor blocking an assimilatory interpretation is evidence that the stop in question is fully released, and not followed by another segment. Gating reveals that before the onset of the stop release, listeners are unable to distinguish a fully released lexical [k], for example, from an unreleased [t] that has assimilated to [k].

[9]There is weak evidence for a difference in the specification of coronals and non-coronals in the pattern of change in priming levels that arises as the stop release gets increasingly truncated, suggesting that for coronals, acceptance as coronal is less dependent on matching to place information.

[10]frequency counts were taken from the Cobuild 17.9 million word corpus

to be accounted for? It could be argued that the different task demands play a significant role. Subjects are clearly able to perceive the phonetic cues in a forced-choice pre-test when they are required to focus specifically on the phonetic detail. When subjects are required to access their mental lexicon rather than make a relative judgement ("A is more [k]-like than [t]-like") however, effects of lateral inhibition (with the more frequent *late* damping down the activation of *lake*) might override the few subtle cues to place of articulation[11].

What *is* clear, from the cross-modal priming experiments on isolated words (Marslen-Wilson and Zwitserlood, 1989; Marslen-Wilson and Gaskell, 1992; Marslen-Wilson, 1993; Marslen-Wilson *et al.*, 1995), is that even very small deviations from the canonical form of a word (i.e. changes of just one phonological feature) will result in a loss of priming. Such loss may be partial or absolute, but in any event *mismatch* reduces the priming effect. A strong representational hypothesis in terms of underspecified features is inadequate for explaining these results. Instead, additional machinery must be postulated, which can relate featural cues and abstract representations to their broader context. We need therefore to look more closely at the specific contextual conditions under which assimilated forms are successful in accessing their source words.

3.5.7 Assimilation in context: match or mismatch?

Using cross-modal repetition priming to examine the perception of assimilated tokens *within sentential contexts*, Gaskell and Marslen-Wilson (1996) compared primes which were phonologically unchanged (place of articulation = coronal) with assimilated primes (lexical coronal changed to labial or velar) in one of two contextual conditions.

In the first experiment, the prime was presented with its lead-in context, but with all speech following the prime removed. Thus the sentential context in (3.3) would be followed by the phonologically unchanged prime *lean*, the phonologically changed prime *leam* or a control such as *brown* or *browm*.

(3.3) *We have a house full of fussy eaters. Sandra will only eat ...*

[11] The responses investigated by Marslen-Wilson *et al.* (1995) were for gates up to but not including the release of the stop that conditions the assimilation context, e.g. the [k] of *cruise* (on the assumption that the word-final stop of the target is unreleased in the context of an abutting stop consonant). This means the only cues to place of articulation are in the formant transitions of the vowel as it goes in to the closure. Although these are shown to be sufficient for success in the forced-choice pre-test, in a lexical identity task they may not be strong enough to fight off lexical effects of frequency.

Targets (e.g. LEAN) were presented visually at the offset of the prime word.

In the absence of following phonological context Gaskell and Marslen-Wilson (1996) found that there was no mismatch effect: the related word primed the target whether or not the final segment was canonical (i.e. coronal) or assimilated (i.e. a fully labial or velar segment). That is, both *lean* and *leam* facilitated responses to LEAN compared with the control. This finding contradicts the standard story of mismatch found from experiments using isolated words (see Sections 3.5.3 and 3.5.6 above), where the slightest change from the canonical form will reduce priming (Marslen-Wilson and Zwitserlood, 1989; Marslen-Wilson and Gaskell, 1992; Connine *et al.*, 1993; Marslen-Wilson *et al.*, 1995). The results of single word cross-modal priming experiments predict that primes involving a full assimilatory change from coronal to labial or velar will produce longer RTs to the target than unchanged primes. Gaskell and Marslen-Wilson found no such effect when primes were heard not in isolation but following a lead-in context (of approximately a sentence and a half in length). Mismatch in these conditions did not reduce the priming effect.

In a second experiment, primes were presented not only with an introductory sentence but also with their following context. The context following the prime was either appropriate for the change in place of articulation of the prime word (= viable context) or inappropriate (= unviable context). For the example in (3.3) above, the viable context would have been *leam bacon* and the unviable context *leam gammon.* The control was *browm loaves.*

It was found that the phonological context following the prime (viable vs. unviable context) had a significant effect on priming. In the viable context, changed primes (i.e. words which involved assimilation) showed no mismatch effect, that is, they showed the same priming effect as unchanged canonical forms. However, in the unviable context, changed primes lead to a reduction in priming. So, while *leam* primed LEAN when followed by *bacon*, there was no facilitation from *leam* in the context of *gammon.*

Gaskell and Marslen-Wilson argue from these results that mismatch matters only when the following context reveals that the change is phonologically unviable. They suggest that subjects, when listening to words in sentential context, apply a process of **phonological inference** to interpret assimilated tokens. In the appropriate phonological context assimilation presents no barrier to priming. Feature-changing variation will not disrupt perceptual interpretation so long as it is viable.

To account for their results, Gaskell and Marslen-Wilson must appeal to a conservative lexical matching process which, though intolerant of mismatch, rejects candidates only when there is unambiguous mismatching information, that is, a process that matches phonetic input onto abstract underspecified representations. If the place node for the coronal undergoing assimilation is unspecified, then a segment articulated at either a labial or velar place will not mismatch the underlying representation. In this case, an assimilated token excerpted from context is as likely to match the appropriate lexical candidate as the unambiguous token.

The process of phonological inference is required to account for the results in the second experiment which involves viable and unviable contexts, since, on a purely representational hypothesis, the deviations involved should match the abstract underspecified representation *regardless* of context. When the word *leam* is heard, the feature [+labial] is attached to the nasal's unspecified place slot by a process of feature spreading, in anticipation of a [+labial] feature to come. As subsequent context arrives, the viability of the change must be assessed. If the following context turns out to be non-labial there will be mismatch, the change is deemed unviable, and there is no facilitation by the unviable prime. Thus highly abstract lexical representations interact on-line with processes of phonological inference to compute the listener's internal representation of the lexical content of the speech stream (Marslen-Wilson *et al.*, 1995).

So, on the basis of the results of cross-modal priming experiments using *isolated* words we would expect tokens involving assimilation to disrupt the process of lexical access (tokens of words involving even a single feature change away from the canonical form have little if any priming effect). However, according to similar experiments involving words *within context*, assimilation in the appropriate (viable) context should not reduce the priming effect.

It should be noted at this point that all the assimilated tokens used by Gaskell and Marslen-Wilson result in English Non-Words. When *lean* assimilates in anticipation of the following labial the resulting output *leam* has no real word competitor. It would be interesting to ascertain whether priming also occurs for assimilated tokens of words like *ran, mat* and *cat*, where the assimilated form presumably activates competitors such as *ram, map* and *cap*.

3.6 Summary

The problem of speech variability has important consequences for the recognition of words. Successful recognition depends on matching information from the acoustic signal to representations stored in the mental lexicon. Current theories involve an initial contact phase, when several lexical hypotheses are contacted, resulting in the activation of multiple candidates, followed by a selection phase during which period one dominant candidate emerges and, when sufficiently more activated than its competitors, 'wins', at which point the word is deemed to have been recognised.

Recognition rates are affected by a word's frequency, and by the size and frequency of the competitor set. Frequent words are recognised more quickly than rare words; words with few competitors are recognised more easily than words which share features with a large number of different lexical items.

TRACE and the Cohort model are two models of spoken word recognition which currently dominate the literature. They differ in a number of respects:

- TRACE involves a process of lateral inhibition (whereby a lexical candidate can reduce the activation levels of its competitors), while the Cohort model depends entirely on natural decay and bottom-up mismatch with no direct mechanism for dealing with competition;
- The Cohort model places great importance on word onsets: in its strictest form, mismatch at the beginning of a word results in failure to access the correct cohort and consequently to recognise the word;
- The architecture of TRACE involves a phonemic level of representation whereas most recent versions of the Cohort model advocate direct access to the lexical representation from featural information.

Priming studies show that while mismatch at word onset generally blocks facilitation, when the competitor environment is sparse rhyme primes will facilitate the target word. Similarly, if a prime that is phonetically ambiguous word-initially has no real word competitor, it will facilitate the target (e.g. *d/task* primes *task*). The importance of matching to onset, then, appears to depend on the density of the competitor environment.

Successful modelling of spoken word recognition requires consideration of both process and representation. Recent advances in phonological theory, in terms of

hierarchical structure and representation, may help models to predict better the results of empirical research on lexical processing. The theory of underspecification in particular has been used to account for listeners' responses in gating tasks to phonetically ambiguous stimuli. However, underspecified representations alone fail to account for the results of priming experiments on words which mismatch features, such as when a word undergoes place of articulation assimilation. In addition to abstract underspecified representations, a process of phonological inference is required, whereby listeners assess the viability of the input *vis-à-vis* its context and the current representation. This requires an *active* listener of the sort advocated by Lindblom (1990a)'s H & H theory.

Chapter 4

Predicting the needs of a Listener

Lindblom's H & H theory claims that the amount of acoustic information provided by the speaker in articulating an utterance depends on the speaker's assessment of their listener's informational requirements: the more predictable the propositional content of the utterance, the less effort a speaker will need to put in to the production. This position represents an overall mapping of the relationship between production on the one hand – the articulatory effort employed by the speaker – and, on the other, a recognition process for both form and meaning.

A major shortcoming in Lindblom's exposition of the H & H theory is the omission of any model of how and what the listener understands. Also omitted is any account of what the speaker *believes* the listener understands. To add flesh to Lindblom's skeletal depiction of the relation between speaker and listener we need to consider the following questions:

- What is it that the speaker is attempting to *achieve* in producing an utterance for her listener? What is her goal, and what, consequently, is her listener's task?
- What information is being communicated, and how is such information *stored* and *accessed*?
- What information can the speaker assume her listener *already knows*?
- *How* does the speaker signal *linguistically* what is informationally important?

This chapter attempts to address these questions. The early part of the chapter focuses on the shared nature of the communicative task: as Lindblom correctly observes, speakers converse in order to share information. The task is essentially

one of establishing a mutually held set of beliefs. The nature of the information represented in a model of discourse is addressed in the subsequent sections. The H & H theory proposes that listeners can use the information represented in this model to help decode the incoming speech signal. Speakers hypo-articulate according to their beliefs about what their listener already knows. The distinction between what is known, or 'Given', and what is 'New' to the discourse is shown to have significant consequences for linguistic production.

4.1 Why speak? (And how to listen?)

I would hope that it is uncontentious to state that a speaker produces an utterance **in order to communicate something** to a listener, and does so by transforming an intention to communicate into a sequence of speech sounds (Levelt, 1989). The listener has then to interpret the incoming speech signal, and access the meaning of the speaker's message. The listener's task, then, is a **decoding** one: he must decode the acoustic signal into a form whereby he can extract meaning. Thus he must translate from the physical to what is ultimately the mental, that is, the cognitive representation of meaning. There are at least three main difficulties the listener has to address, therefore:

- The first is the problem of **lexical access**, that is, converting the speech stream into a string of words, with their associated meanings. This process was discussed in detail in the preceding chapter.
- Secondly, these individual words must be combined to form a **propositional representation of meaning**: a semantics for the utterance as a whole.[1] (Since the work of this thesis is not directly concerned with the process of constructing the propositional meaning of an utterance from its componential elements (words), the process is not discussed further here.)
- The third problem is knowing how to attach the accessed lexical information to the stored mental representation of the current discourse, i.e. **situating the present contribution**.

According to Lindblom, the information available for the first process – the recognition of words – is dependent upon the speaker's view of the current status of

[1] Whilst this process will necessarily be initiated after the onset of lexical access, it does not depend on all words in the utterance being successfully recognised before starting.

the listener's mental representation of the discourse. In order therefore to account for what is or is not available in the acoustic stream, we need to consider first the representation of information within a model of discourse. It is to this issue that we now turn.

4.2 Building a mental model of discourse

How is knowledge of the world organised in human memory and how is it activated in the process of discourse understanding? Various researchers have attempted to shed some light on these questions, and in doing so, have introduced a bewildering number of different terms: Frames, Scripts, Scenarios, Schemata (see Brown and Yule, 1983, for details and references). Despite the variation in terminology, however, the analyses have much in common: they all explore the way in which discourse is interpreted via an appeal to stereotypic knowledge.

Johnson-Laird (1983, 1987) takes one step further: he uses the term **mental model** to refer to the 'models' of reality that humans construct in order not just to understand discourse but to reason about the world in general. Mental models are cognitive representations of the way the world is, or at least, the way it is perceived to be. Influenced by model theoretic semantics, Johnson-Laird has argued that we can construct and manipulate models of both the real and the imaginary; models are representations of possible states of affairs at a particular point in time and space. That is, they are constructed locally – relative to the context of utterance – and vary from one individual to another.

Discourse comprehension, under this view, depends on three levels of representation: the **phonemic**, the **propositional** and the **mental model**. The phonemic level represents the decoding of the incoming acoustic signal into language-specific segmental chunks by which access is made to stored lexical information. The propositional level combines the accessed meanings of the individual lexical items into a (truth conditional) semantics of the utterance as a whole. The mental model incorporates the meaning of the utterance into a larger framework of previously stored information. The mental model is constructed on the basis of the truth conditions of the propositions expressed by the sentences in the discourse, which in turn depend on the context of utterance and inferences from background knowledge. It is, of course, this 'context of utterance' and 'background knowledge' to which Lindblom is referring when he talks of 'sources of information' available to the listener to help him decode the incoming signal.

The importance of a cognitive model of discourse for understanding the meaning of an utterance is emphasised by Givón, who states that: "comprehension is *synonymous* with the construction of a structured mental representation of the text" (Givón, 1995, page 64, emphasis mine). Givón (with others) distinguishes between two different mental 'text-traces':

- the **working memory buffer**, with its severely limited capacity of roughly 8 - 20 seconds of verbatim text, and
- **episodic memory**, flexible and context-dependent, in which text is represented as a **network of connected nodes**. Such networks demonstrate both **hierarchical** structure (clauses[2] form chains which form paragraphs) and **sequentiality** (nodes have connections to preceding and following adjacent nodes).

Given the constraints imposed by the working memory buffer, the main on-line task of language processing is to "determine where and how to attach the new information in the clause in the episodic mental representation of the current text" (Givón, 1995, page 103).

If we assume, with these and other authors, that the process of communication involves the construction by speakers and listeners of a mental representation of the discourse, then we can suppose that this model makes available to the listener the information that Lindblom proposes speakers consider in assessing the listener's needs.

4.3 Establishing coherence: a collaborative enterprise

The mentally represented text, and the mental processes that partake in constructing that mental representation, together create what researchers refer to as **coherence**. The problem of establishing coherence has been characterised as the creation of order out of chaos (see Tannen, 1984), of imposing structure to unify what on the surface might look like a jumble of disjointed utterances. Grice (1975)'s maxim of Relevance is pertinent here: so long as listeners assume that the speaker's contribution is relevant they will attempt to interpret the utterance

[2] A clause is defined as "the minimal unit for accruing new language-coded information into episodic memory" (Givón, 1995, page 97)

in relation to what has (immediately) preceded, even if, on the surface, there appears to be no direct connection, as in (4.1).

(4.1) a. *We're running out of petrol!*

b. *The Savacentre's downhill from here.*

Successful interpretation of the second contribution (4.1b) depends either on previous knowledge or on the ability to infer (based on the assumption that the utterance is relevant) that the Savacentre sells petrol.

For (4.1) to be a coherent text, the two interactants need to *align their mental discourse representations.* They do this by establishing what they mutually believe to be true. For the example above, both interactants should believe that:

- more petrol is required (urgently)
- the Savacentre sells petrol
- the Savacentre is at the bottom of the hill
- the Savacentre is open
- there is sufficient petrol left to get as far as the Savacentre

In this way utterance (4.1b) is seen to 'fit in' or *coordinate* with the preceding utterance (4.1a). Incoming discourse contributions are attached to the current representation, in such a way that they fit in – or *cohere* – with what's gone before.

According to Gernsbacher and Givón (1995), coherence is a property which *emerges* during speech production and comprehension as the result of a 'collaborative process' involving two minds attempting to achieve, simultaneously, a number of goals. Interlocutors work together to establish a **mutually agreed, shared representation of the discourse**: a representation of the entities to which they have referred, of the relations that hold between them, and the properties associated with each. Gernsbacher and Givón claim that during conversation the negotiation takes place collaboratively between two or more active participants, and that coherence emerges not in the text itself but in the two collaborating minds.

This interpretation of coherence is echoed by Wilkes-Gibbs (1995), who proposes that coherence be viewed as **coordination**, with the solution to the problem

being a collective rather than individual achievement. Clark shares this view of language as a joint, collaborative activity (Clark and Schaefer, 1987a; Clark and Brennan, 1991; Clark, 1992, 1996). Clark likens language activity to the dancing of a waltz, where the dance amounts to more than the sum of the constituent parts. Clark distinguishes **participatory** from **autonomous** acts, and argues that participatory acts – those acts used in joint actions such as language or waltzing – only make sense in the context of the joint action of which they form a part. When one participant engages in their part alone there is a difference in *intention* which is manifest in the way actions are **coordinated**: when dancing together, partners are expected to coordinate their movements to each other. If dancers fail to accommodate in this fashion then they cannot really be said to be dancing *together*.

Clark argues that language operates along similar principles of coordinated activity. For Clark, the basis of conversation is **common ground**. We can only base a conversation on what we share, and the goal of any conversation is to *extend* what we share, and thus add to our common ground. We therefore tailor our utterances to the (perceived) common ground of our fellow participants. While Clark is talking here about tailoring the propositional content, rather than the acoustic output, there is a clear parallel between his and Lindblom's position: both claim that we alter our linguistic behaviour according to our beliefs about the knowledge state of our listener.

Speakers' ability to accommodate their contributions to the perceived common ground of their listener is demonstrated in a study by Clark and Schaefer (1987b) where participants were asked to *conceal* the information they were exchanging from an overhearer. Pairs of friends had to identify a number of well-known local landmarks to each other, without revealing this information to an overhearer, who was unknown to either member of the pair. They achieved this successfully, by appealing to what Clark and Schaefer refer to as **private common ground**: knowledge of an event or occasion that would be shared by only the few people directly involved. In a control condition, where interactants were not aware of any overhearer and there was therefore no requirement for concealment, the pair used references which would have been familiar to any local resident, that is, they made use of more general, rather than private, common ground. Clark concludes that this adjustment in strategy reflects the fact that speakers are aware of common ground, and are able to accommodate their use of it appropriately.

Sensitivity to common ground suggests that speakers ought to be aware of the pitfalls in attempting to extend shared knowledge: it would be foolhardy to assume that an utterance has automatically been heard and understood. Consequently, Clark and Schaefer (1987a) claim that common ground is added to through the collaborative process of **contributing**. A contribution involves two phases: a **presentation** phase, in which A presents an utterance for B to hear, and an **acceptance** phase, in which B provides evidence of understanding. Only once a presentation has been accepted is it deemed to have been **grounded**, that is, added to the interactants' common ground (Clark and Schaefer, 1987a). One of the problems with Clark's presentation/acceptance model of grounding is that the process is iterative: it will continue until sufficient evidence has been given in the acceptance phase for the original speaker to consider understanding to have been achieved. But if each acceptance is, itself, a presentation, which therefore requires an acceptance, how can the system 'bottom out' and end?

Recently, Heeman and Hirst (1995) have presented a computational model of collaborating which revises Clark's presentation/acceptance process to include three moves: **present**, **judge**, and **refashion**. As Davies (1997) observes:

> *"The idea of acceptance is changed from being a move in the process to being a boolean value which controls the process:*
>
> ***present***
> ***judge**(judgement)*
> *while (judgement* $\neq$ ***accept**)*
> ***refashion***
> ***judge***
> *end-while*
>
> *Adapted from Heeman and Hirst (1995:354)*
>
> *(Davies, 1997, page 29)*

This makes it clear how the acceptance phase should terminate: refashioning continues until the presentation is judged to have been accepted. Heeman and Hirst also emphasise that *both* participants have access to the judging and refashioning process.

Mutual knowledge, then, requires a process of acceptance, and not just the broadcasting of information. Both partners in a dialogue work together, engaging in turn-taking and question asking, to establish common ground and thus, a mutual representation of the discourse (Wilkes-Gibbs, 1995). Meaning in communication is therefore a *social* construction:

> *"A coherent discourse, or a coherent utterance within one, develops from the moment-by-moment activities of the participants, working together in regular ways to produce evidence of a shared understanding."*
> *(Wilkes-Gibbs, 1995, page 241)*

This view of coherence as a collaborative process based on negotiation is a positive and optimistic view of language behaviour which assumes that dialogue participants are cooperative and considerate of each others needs. Just how cooperative conversation partners really are is a question that will be raised and discussed a little later (see Section 4.9); for the present it will be sufficient to assume that speakers create coherence in discourse by establishing a similar mental representation of the current text (whether written or spoken), and that they achieve this through negotiation.

4.4 What to communicate: uttering information

Clearly, then, one major step in constructing an account of conveyed information is relating incoming information to previously stored information. As will be shown below, the listener's interpretation of the propositional representation, and indeed the choices made by the speaker in relation to what she talks about and how, is affected, in particular, by what the speaker and listener each believe the other already knows about the discourse. The structuring – or packaging – of propositional information, then, depends on what the interlocutors believe to be relevant or salient in a particular context. Vallduví (to appear) defines information packaging as follows:

> *"A sentence, in one of its facets, may be viewed as a structural vehicle used to transfer some piece of knowledge (a proposition) from speaker to hearer. Information packaging is the speaker's tailoring of this structural vehicle to suit some 'communicative' aspect of the transfer of knowledge (propositional content) to the hearer. In other words, when communicating a proposition* p *a given speaker may encode* p *in different sentential structures according to his/her beliefs about the hearer's knowledge and attentional state with respect to* p."
> *(Vallduví, to appear, page 2)*

In other words, information structure is *context dependent*; the propositional content of an utterance is interpreted in the light of the current mental representation

of the discourse available to speaker and listener. What kind of information, then, is stored in the discourse representation, and how is it accessed?

4.4.1 Utterance as instruction to update the mental model

Prince (1981) proposes that text be viewed as: "a set of instructions from a speaker to a hearer on how to construct a particular DISCOURSE MODEL" (1981: 235). She continues:

> *"The model will contain* DISCOURSE ENTITIES, ATTRIBUTES, *and* LINKS *between entities. A discourse entity ... [or]* DISCOURSE REFERENT *... may represent an individual (existent in the real world or not), a class of individuals, an exemplar, a substance, a concept* etc." *(*ibid.*)*

Discourse entities, she says, may be thought of as *hooks* on which to hang attributes. A similar view is reflected in Givón (1995)'s description of the cognitive representation of discourse, where a network of connected nodes is accessed via a clause's *topical referent.* The referential NP acts as a node label, or filing address for the clause. Thus, a mental representation of discourse consists of a network of connected nodes, where each node is associated with a referring expression that links the node to an entity. The nodes themselves contain information relating to that entity: the properties associated with it, the relations that hold between it and other entities *etc.* Connections between nodes represent links between different referring expressions.

In uttering a declarative sentence, a speaker specifies items of information which she believes her hearer will be able to **access** from his representation of the discourse, and then specifies, further, what properties the hearer ought to **assign** to these items, and/or modifications the items should undergo (Hajičová, 1991).

The notion of referent accessibility is discussed in Section 4.6.1. But before considering how referential information might be stored and accessed in a cognitive representation of text, we turn to the distinction, implicit from the preceding discussion, between 'items of information already known about' and 'new bits of information' to be incorporated in to the current representation. This is a crucial division, which has been variously described as a distinction between 'Given' and 'New', 'Topic' and 'Focus', or 'Theme' and 'Rheme' by a number of researchers who have used these different terms with varying degrees of overlap in meaning and context. In the following sections I present some of the interpretations that

have been made of what I shall refer to as the 'Given/New' distinction. I then consider how this distinction might be represented within a cognitive model of discourse, before looking at the linguistic evidence that reflects speakers' usage of the distinction.

4.5 Distinguishing 'Given' from 'New'

Prince (1981) observes that the objective information conveyed in an utterance is not conveyed "on a single plane", but that there is an **informational asymmetry**: some units seem to convey or represent "older" information than others. The crucial factor, she suggests, is that speakers tailor an utterance to meet the needs of the intended receiver, they decide how to package up the information they wish to convey, according to what they believe their listener already knows:

> *"Information-packaging in natural language reflects the sender's hypotheses about the receiver's assumptions and beliefs and strategies." (Prince, 1981, page 224)*

As Prince (1981) illustrates, the terms **Given** and **New** have been applied with varying interpretations by different writers. Prince distinguishes at least three different levels of Givenness:

- **Givenness$_p$: Predictability/Recoverability**
  associated with: Halliday, Kuno (e.g., Halliday, 1967; Kuno, 1974)
  "The speaker assumes that the hearer CAN PREDICT OR COULD HAVE PREDICTED that a PARTICULAR LINGUISTIC ITEM will or would occur in a particular position WITHIN A SENTENCE."

- **Givenness$_s$: Saliency**
  associated with: Chafe (e.g., Chafe, 1976)
  "The speaker assumes that the hearer has or could appropriately have some particular thing/entity/ . . . in his/her CONSCIOUSNESS at the time of hearing the utterance."

- **Givenness$_k$: 'Shared Knowledge'**
  associated with: Clark (e.g., Clark and Haviland, 1977)
  "The speaker assumes that the hearer 'knows', assumes, or can infer a particular thing (but is not necessarily thinking about it)."

All three levels involve cases, says Prince, where some item is Given for extralinguistic reasons, but they vary in their analyses of when and how such items can be considered Given.

4.5.1 Givenness$_p$: Predictability/Recoverability

For Kuno, "An element of a sentence represents old, predictable information if it is recoverable from the preceding context; if it is not recoverable, it represents new, unpredictable information" (Kuno, 1978, pages 282-3). In Kuno's interpretation recoverability equates to 'deletability': whether or not a particular linguistic item is sufficiently predictable to undergo ellipsis. In particular, Kuno was concerned with the relation between deletable pronouns and their antecedents. However, as Prince (1981) indicates, this approach forces the two pronouns in (4.2a) and (4.2b) to be interpreted differently, when, intuitively, one might want to treat both of them as Given. Clearly Kuno misses the syntactic constraint.

(4.2) a. ***John***$_i$ *paid Mary and [**he**$_i$] bought himself*$_i$ *a new coat.*
(deletable)

b. *Mary paid **John***$_i$ *and **he***$_i$ *bought himself*$_i$ *a new coat.*
(non-deletable)

Halliday defines Given/New quite differently, but again, his criterion is specifically *linguistic*. New information is differentiated from Given information by **intonation**: New information is identified as intonationally marked or unmarked focus (see Section 4.8 for a more detailed discussion of Halliday's position). New information is said to be focal "not in the sense that it cannot have been previously mentioned, although it is often the case that it has not been, but in the sense that the speaker presents it as not being recoverable from the preceding discourse" (Halliday, 1967, page 204). For Halliday it is the *speaker* who determines what is or is not to be treated as New: Givenness, under this approach, is not an independent property of the text itself but a reflection of the speaker's interpretation of the text.

4.5.2 Givenness$_s$: Saliency

Givenness$_s$ refers to what the speaker believes her hearer to be **conscious** of at the time of utterance. According to Chafe (1976)'s notion of Given/New – which

he takes to be a binary distinction – known[3] items that are introduced into the discourse for the first time are as New as unknown ones. The bold-faced NPs in (4.3a) and (4.3b) (from Prince, 1981, page 229) are therefore equally New.

(4.3) a. *I saw* ***your father*** *yesterday.*

b. *I saw* ***a two-headed man*** *yesterday.*

For an NP to qualify as Given, in Chafe's use of the term, its referent must have been *explicitly* introduced in the text or else be present in the physical context; inferentially related NPs cannot be Given (unless the inference is one of categorisation). Thus, in (4.4) (originally from Haviland and Clark (1974)) Chafe would interpret *the beer* as Given in (4.4a) and New in (4.4b).

(4.4) a. *We got some beer out of the trunk.* ***The beer*** *was warm.*

b. *We got some picnic supplies out of the trunk.* ***The beer*** *was warm.*

4.5.3 Givenness$_k$: 'Shared Knowledge'

In Clark and Haviland's analysis of Given/New, Given information is "information [the speaker] believes the listener already knows and accepts as true" (Clark and Haviland, 1977, page 4). It is immaterial whether the hearer knows the information by having been told about it explicitly, or by indirect means of inference. In the example above, *the beer* is Given$_k$ in both (4.4a) and (4.4b).

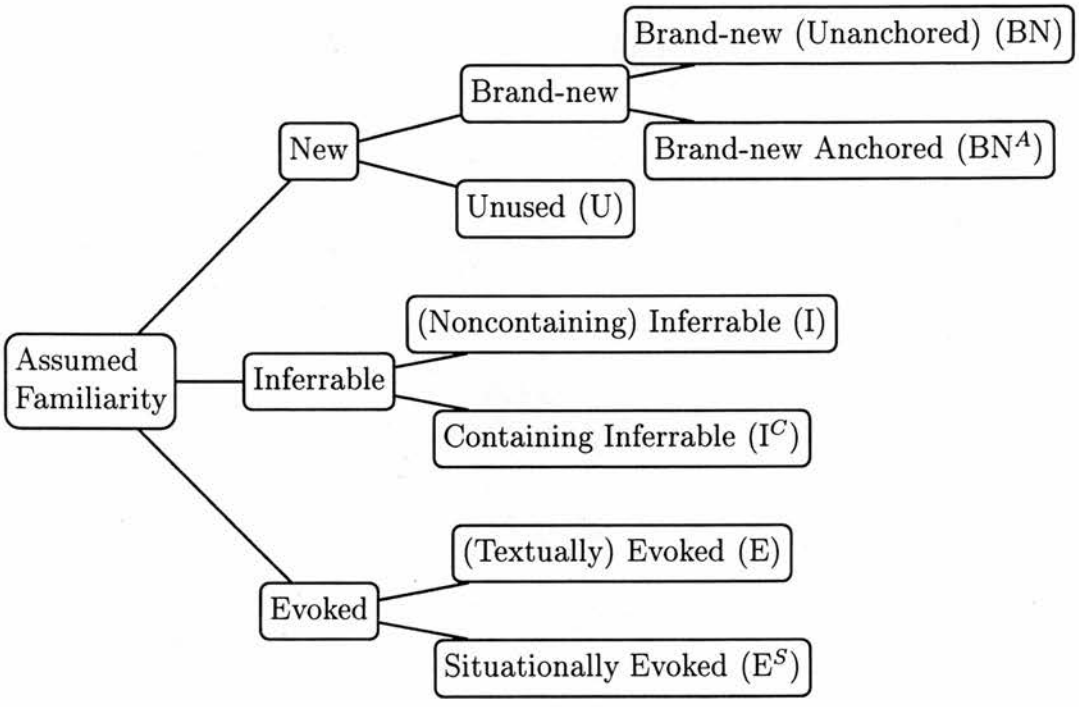
4.5.4 Givenness as 'Assumed Familiarity'

Clearly, the three interpretations of Givenness described above are related. The writers are agreed that, at some level, Givenness relates to "what the speaker thinks is or should be or could appropriately be in the hearer's mind" (Prince, 1981, page 232). In other words, Givenness is associated with what the speaker assumes is already stored in the hearer's mental representation of the discourse, or, in the case of Givenness$_k$, what the hearer can access from generic-lexical knowledge, to use Givón (1995)'s term.[4]

[3] in the sense of "background knowledge" i.e. independent of what has been said

[4] It should be noted here that interest in what an individual knows or hypothesises about another individual's belief state is restricted to such knowledge and hypotheses that affect the forms and understanding of **linguistic** productions.

Prince goes further. She takes Clark and Haviland (1977)'s interpretation of Givenness$_k$, she discards the term 'Shared Knowledge' in favour of **Assumed Familiarity**, and presents a taxonomy of values of Assumed Familiarity, which involves a three-way (rather than binary) distinction between **New**, **Inferrable** and **Evoked** entities in discourse. Further subdivisions result in the following set of relations:



These distinctions can best be illustrated with reference to some examples.

(4.5) a. *I took* ***a taxi*** *this morning and* ***the driver*** *hadn't a clue where* ***he*** *was going.*

b. ***A guy I work with*** *says* ***he*** *knows* ***Mike.***

c. *I've heard* ***the Duke of Edinburgh*** *is opening the new Informatics building.*

d. *Excuse* ***me,*** *do* ***you*** *have change of a fiver?*

e. *It looks as if* ***one of these bananas*** *is bruised.*

Brand-new refers to entities that have to be created from scratch by the hearer, such as *a taxi*, or *a guy I work with* in (4.5a) and (4.5b) above. Brand-new entities may or may not be **Anchored**. An entity is Anchored if "the NP representing it is LINKED, by means of another NP, or 'Anchor', properly contained in it, to

some other discourse entity" (Prince, 1981, page 236). The NP *a guy I work with* is Anchored since the discourse entity that the hearer creates for this reference will immediately be linked to his discourse entity for the speaker (the person with whom the guy works).

Unused entities are those which are new with respect to the discourse, but which the speaker believes her hearer already knows, and has a corresponding referent for in long term semantic memory. The NPs *the Duke of Edinburgh*, and *Mike* in (4.5c) and (4.5b), respectively, are examples of Unused entities.

Evoked refers to entities which already exist in the discourse model. These entities will have been evoked by the hearer in one of two ways. Either the speaker will have made reference to the entity earlier in the discourse, in which case it is **Textually Evoked** (or simply Evoked), or alternatively the hearer will have evoked the entity for reasons associated with the discourse situation, but independently of what has been said. **Situationally Evoked** entities include discourse participants and salient features of the 'extratextual' context. In examples (4.5a) and (4.5b) above, *he* is Evoked, while, in (4.5d), both *me* and *you* are Situationally Evoked.

According to Prince, **Inferrables** are the most complicated type of discourse entity. "A discourse entity is Inferrable if the speaker assumes the hearer can infer it, via logical – or, more commonly, plausible – reasoning, from discourse entities already Evoked or from other Inferrables" (Prince, 1981, page 236). In (4.5a), *the driver* is Inferrable from the Evoked entity *a taxi*, via knowledge that taxis are vehicles, and vehicles have drivers. Similarly, *the beer*, in (4.4b) above, is Inferrable from *some picnic supplies*, since picnic supplies regularly include various foodstuffs – bread, cheese, salad, crisps – and something worthwhile to drink. Prince also introduces a special subclass of Inferrables, called **Containing Inferrables**. These are Inferrable NPs which contain, within them, the reference from which they can be Inferred. In (4.5e), *one of these bananas* is a Containing Inferrable, as the bruised banana (*one of* ...) is inferrable, by a set-member inference, from the whole bunch, (*these bananas*), which, themselves, are Situationally Evoked by being physically present at the time of utterance.

From an empirical analysis of several naturally occurring texts, Prince (1981) concludes that there is a preferred hierarchy for what type of entity is used by speakers in a discourse. Drawing up the following Familiarity scale, she observes that there is a tendency to use an NP that is *as high on the Familiarity Scale as felicitously possible.*

Prince's Familiarity Scale:

$$\textit{(top)} \quad \left\{ \begin{array}{c} \mathrm{E} \\ \mathrm{E}^{S} \end{array} \right\} > \mathrm{U} > \mathrm{I} > \mathrm{I}^{C} > \mathrm{BN}^{A} > \mathrm{BN} \quad \textit{(bottom)}$$

Presented with the set of choices in (4.6), a speaker will prefer (and be expected by the hearer) to use an NP as near to the top of the list as she can, consistent with what she believes her listener to know about the intended referent. Using an NP which is lower down the Familiarity scale than is deemed appropriate with respect to these beliefs would be interpreted, if found out, as deviant (deliberately evasive, childish, or whatever).

(4.6) a. ***I** have a daughter called Helen.* (E^{S})

b. ***Cathy** has a daughter called Helen.* (U)

c. ***One of the researchers who works at the HCRC** has a daughter called Helen.* (I^{C})

d. ***A researcher who works at the HCRC** has a daughter called Helen.* (BN^{A})

e. ***A researcher** has a daughter called Helen.* (BN)

If the speaker referred to someone as *one of the researchers who works at the HCRC* when she believed her addressee to know that someone by name (*Cathy*), then failing to use the Proper Name, *Cathy*, would signal that the speaker was attempting to mislead her listener, or at least failing to be cooperative. In Prince's words:

> *" The use of an NP representing a certain point of the scale implicates that the speaker could not have felicitously referred to the same entity by another NP higher on the same. The recognition of such a scale permits this sort of implicature to be subsumed under the Gricean maxim of Quantity."*
> *(Prince, 1981, page 245)*

It would appear, then, that speakers distinguish between information that they consider to be 'New' to their listener, and information that is deemed 'Given' – **knowable** by the listener – either through evocation (textually or situationally evoked) or inference. Further, a speaker's choice of **referring expression** seems to relate to the degree to which the referent is believed, by the speaker, to be Given for her listener. That is, referential Givenness reflects a speaker's assumptions

about the **cognitive status** of a referent in the mind of her addressee. The next section considers how Givenness might be represented in a mental model of discourse; it introduces the concept of **Accessibility** and considers the effects of accessibility on a speaker's choice of referring expression.

4.6 The cognitive representation of Givenness

Speakers vary the way they refer to entities according to what they believe their hearer already knows about the referent, i.e. the degree to which the referent is hypothesised to be Given (Chafe, 1976; Clark and Haviland, 1977; Prince, 1981; Ariel, 1990, 1991; Gundel *et al.*, 1993; Lambrecht, 1994).

Reference to a (Brand-)New entity is, essentially, an instruction to the listener to **create a new discourse referent** in their model of discourse (Prince, 1981). Reference to a non-(Brand-)New entity, on the other hand, requires the listener to access an **existing** representation, and attach to this, a piece of additional information (Hajičová, 1991). In other words, an entity can be treated as Given in the referential sense if the speaker can assume that her listener has a representation of this entity in long or short term memory.

Givón (1995), it may be recalled (see Section 4.2), presents a mental model of discourse which consists in a network of connected nodes, each of which is accessed via a topical referent. The connections between nodes represent links between different referring expressions. This connectivity of a node to other nodes in the network is referred to by Givón as **grounding**[5]. According to Givón, connections to nodes in a network can be made in one of two main directions:

- **Cataphoric grounding** involves the opening of pending connections in yet-to-be completed structure, in *anticipation* of text that is in the process of being constructed;
- **Anaphoric grounding** involves connections 'backwards' to earlier parts of structure; the referent is assumed to be represented in – and can be retrieved from – some pre-existing mental structure in the hearer's mind.

Note that the 'pre-existing mental structure' is not restricted to a structure of the current discourse, thus it includes reference to structures stored in Encyclopaedic

[5]Note this is a different grounding process from Clark and Schaefer (1989)'s **content grounding**, which is an explicitly *social* process (see Wilkes-Gibbs, 1995).

(Ariel, 1990) or Generic-Lexical (Givón, 1995) Knowledge. This allows Givón to include Prince's Inferrables, and, indeed, Unused entities, as anaphorically groundable.

Cataphoric grounding has received relatively little attention in the literature, claims Givón (1995), and is discussed almost exclusively in terms of indefinite reference and the *absence* of anaphoric grounding. Givón illustrates a variety of grammar-cued devices to ground a referent cataphorically, drawing on the distinction between NPs which are likely to be salient and those which will not be relevant:

> *"Referent NPs are identified as either those that will be important, topical, and thus persistent in the subsequent discourse, or those that will be unimportant, non-topical, and thus non-persistent. Topical referents are most commonly given special grammatical marking." (Givón, 1995, page 65)*

Anaphoric grounding has been much more extensively studied. It serves to establish "a mental connection between the referent's occurrence in the current text-location and its previous anaphoric trace in some extant mental structure" (Givón, 1995, page 68). Ariel (1990) uses the term **context-retrieval expression** to cover all expressions, including but not exclusively NPs, which cause listeners to access previously stored representations. We might ask, then, how it is that listeners know which discourse referent to access on hearing a context-retrieving expression?

Givón distinguishes three types of mental structures used to ground re-introduced referents anaphorically:

- model of the current speech situation
- model of permanent generic-lexical knowledge
- episodic model of the current text

Referents (or other coherence elements) are grounded to the **current speech situation** via anaphoric devices marking spatial or temporal relations to the discourse participants, such as the use of the pronouns *I*, *you*, *we*; location markers like *this one*, *that one*, *here*, and *over there*; and time markers such as *now*, *then*, *tomorrow*, *last week*. These referents correspond to Prince (1981)'s category of Situationally Evoked.

Grounding to **generic-lexical knowledge** – the culturally shared knowledge in permanent semantic memory – is achieved in two ways. Some referents are 'globally accessible', that is, uniquely identifiable at any time by all members of the relevant speech community (nation-state, religion, town, family). Such referents might include *the sun*, *the prime-minister*, *the castle*, *Auntie Rosie*, and correspond with Prince's category of Unused. Most referents however will be grounded to generic-lexical knowledge via a combination of connections to generic-lexical knowledge and episodic text-based access. This 'double grounding', as Givón calls it, has been referred to elsewhere as **frame-based** or **script-based** reference (Anderson *et al.*, 1983; Walker and Yekovich, 1987). In (4.7) the definite referent *the bus* receives its anaphoric grounding partly from the antecedent referent *school* in the preceding text, and partly from generic-lexical knowledge of the frame 'school' and its sub-component 'bus'.

(4.7) *My daughter was late for* ***school*** *today, she missed* ***the bus***

Frame-based referential access is often accomplished through conventional knowledge of relations like 'whole-part', 'possessor-possessed'. This is similar to Haviland and Clark (1974)'s example about picnics and beer, i.e. Prince (1981)'s category of Inferrables.

A huge variety of devices is available for grounding referents to anaphoric traces in the **episodic model of the current text**, that is, grounding to Textually Evoked references (Prince, 1981). The choice of device – zero anaphora, unstressed pronoun, stressed pronoun, full lexical noun, Left-dislocated NP *etc.* – tends to indicate the size of anaphoric gap between referent and antecedent, which, in turn, will reflect the ease or difficulty with which the antecedent is likely to be accessed (see Gundel *et al.*, 1993).

4.6.1 Accessibility: retrieving antecedents

Not all mental representations are equally accessible to addressees at any given moment in the discourse; they are accessible in *varying degrees*. It has been suggested (Ariel, 1990, 1991) that context-retrieving expressions – including referring expressions – are in fact **Accessibility markers**. In other words, a speaker chooses between various referring expressions in order to mark accessibility differences for her listener's convenience. It is the *degree of accessibility* that then guides referent retrieval, rather than the contextual source – the linguistic detail – itself. According to Ariel: "addressees are guided in antecedent retrievals by

considering the degree of Accessibility signalled by the marker" (Ariel, 1991, page 444).

Accessibility reflects a **depth of storage problem**. The distance between the most recent previous mention of the antecedent and the current NP will crucially affect the entity's level of activation and therefore its accessibility. Distance between last and current mention is not simply a matter of number of intervening words or clauses, claims Ariel (1991), but is related to frame, or 'paragraph' structure. Ariel refers to this as the **Unity Criterion**.[6]

There are two other important features that affect accessibility:

- **prominence:** the salience, or importance, of a person or event to both hearer and speaker. A close member of the family will be more accessible than an infrequently met friend[7]; similarly, personal involvement might make the referent *Lockerbie*, *Hillsborough*, or *Dunblane*, for example, more accessible than it would be to someone totally unconnected with these tragic events.

- **competition:** the number of other candidates that can potentially serve as antecedents to the current NP. It has been observed (e.g., Clancy, 1980) that as intervening NPs accumulate, the number of zero-anaphors and pronouns gradually reduces until they disappear from usage.

Ariel (1991)'s linguistic codification of accessibility is based on the following three criteria:

informativity the amount of lexical information contained within the context-retrieving expression;

[6]Walker (1996) discusses this storage issue in some depth, contrasting the different ways in which researchers have characterised what she calls the *limited attention constraint*: the restrictions placed on a listener's attention span which constrain the selection of potential candidates in the resolution of pronominal anaphora, ellipsis *etc.* Whilst Clark and Sengul (1979), for example, have advocated a constraint of *linear recency*, others, such as Grosz and Sidner (1986), talk in terms of *hierarchical recency*. Walker (1996) reviews Grosz and Sidner (1986)'s STACK model of attentional state, observing that it includes no constraints on the length, depth or amount of processing required for an embedded segment. She illustrates how the STACK model fails to account for the data in her corpus of Informationally Redundant Utterances (IRUs), and offers an alternative *cache* model of attentional state. The CACHE is a limited capacity, almost instantaneously accessible, memory store, representing *working memory*. It uses store and retrieve operations to process hierarchically structured discourse intentions. Whilst all discourse conversants maintain their own cache, some conversational processes are necessarily dedicated to keeping these caches synchronised.

[7]This becomes apparent when your daughter shares the same name as a long-standing college friend, and you find yourself accessing the wrong 'index card' when talking with a second, mutual, friend, prompting a query of "which Helen are we talking about, here?".

rigidity the level of ambiguity (how uniquely identifying the expression is);

attenuation the physical characteristics of production, i.e. stress, duration, vowel reduction, assimilation *etc.*

The more informative, more rigid and less attenuated an expression, the *lower* the accessibility rating or score.

Ariel (1991) extends the mechanism of discourse accessibility marking to the distribution of sentential anaphoric expressions. In an analysis of zero/pronoun choices in Hebrew, she demonstrates how the richly informative, rigid and fully articulated 1st and 2nd person pronouns mark relatively low accessibility, while the present tense inflection, which is uninformative, ambiguous, and attenuated, marks an extremely high degree of accessibility. Intermediate levels of accessibility are marked by a variety of forms such as cliticized pronouns, past and future tense person inflection *etc.* which form a hierarchy of accessibility. Ariel illustrates how a shift in discourse topic will prompt a change from cliticized to full pronoun, for example. She observes that:

> *"speakers make consistent choices in their anaphoric expressions, favouring relatively higher Accessibility markers when the antecedent is highly available (when it is a topic ...* etc.*)"*
> *(Ariel, 1991, page 462).*

Accessibility, then, reflects the cognitive status of a referent, with different forms of context retrieving expressions being "specialised for a specific degree of memory Accessibility" (Ariel, 1991, page 462). In other words, the cognitive concept of accessibility has a linguistic correlate: the form of referring expression.

4.6.2 Cognitive status and forms of referring expression

The relation between different forms of referring and the cognitive status of the referent is explored by Gundel *et al.* (1993) who propose a hierarchy of six cognitive statuses – the **Givenness Hierarchy** – which relates to the conventional meanings signalled by determiners and pronouns. Gundel *et al.* looked at the distribution and interpretation of referring expressions in naturally occurring discourse across five languages – English, Japanese, Mandarin Chinese, Russian, and Spanish – and found universal evidence for a correlation between linguistic form and cognitive status. They argue that different determiners and pronominal forms conventionally signal different cognitive statuses and thereby enable addressees to

restrict the set of possible referents. Variation in cognitive status reflects differences in information about the location in memory of the appropriate antecedent and the attention state of the hearer.

The Givenness Hierarchy: (Gundel *et al.*, 1993)

in focus	>	activated	>	familar	>	uniquely identifiable	>	referential	>	type identifiable
{*it*}		{ *that* / *this* / *this* N }		{*that* N}		{*the* N}		{indef *this* N}		{*a* N}

The hierarchy involves a unidirectional entailment, such that a position on the hierarchy has all of the statuses to its right; thus a reference that is 'familiar' is also 'uniquely identifiable', 'referential', and 'type identifiable'. The individual statuses are characterised as:

Type Identifiable The speaker assumes the addressee is able to access a representation of the *class of objects* described by the expression; necessary for any nominal expression; sufficient for the use of the indefinite article *a* in English.

(4.8) *I couldn't sleep last night.* ***A dog*** *(next door) kept me awake.*

The NP *a dog* in (4.8) is appropriate if the addressee can be assumed to know the meaning of the word *dog* and therefore can understand what type of thing the phrase *a dog* describes.

Referential An expression is used referentially if the speaker intends it to refer to a *particular object or objects*; necessary for the appropriate use of all definite expressions; necessary and sufficient for indefinite *this* in colloquial English.

(4.9) *I couldn't sleep last night.* ***This dog*** *(next door) kept me awake.*

The addressee must either retrieve an existing representation of the intended referent or, more usually, construct a new representation.

Uniquely Identifiable The speaker assumes the addressee can identify the referent on the basis of the nominal alone; necessary for all definite reference; necessary and sufficient for use of the definite article *the*.

(4.10) *I couldn't sleep last night.* ***The dog*** *(next door) kept me awake.*

The addressee will usually be expected to have an existing representation for the referent, but identifiability does not have to be based on previous familiarity as long as there is enough descriptive content encoded in the nominal itself. In (4.10) the material in parentheses would provide sufficient information for the reference to be felicitous, even if the addressee had not known previously that the speaker's neighbour has a dog.

Familiar The speaker assumes the addressee already has a representation of the referent (in either long or short term memory); necessary for all personal pronouns and definite demonstratives; sufficient for use of the demonstrative determiner *that.*

(4.11) *I couldn't sleep last night.* ***That dog*** *(next door) kept me awake.*

In (4.11) the addressee must already know that the neighbour has a dog for *that dog* to be felicitous.

Activated The speaker assumes the addressee has a representation of the referent in short term memory (cf. Chafe (1976)'s view of Givenness; see also Hajičová (1991)); necessary for the appropriate use of all pronominal forms and the demonstrative pronoun *this*; sufficient for use of the demonstrative pronoun *that.*

(4.12) *I couldn't sleep last night.* ***That*** *kept me awake.*

In (4.12) the pronoun *that* is only appropriate if, during the speech event itself, a dog has been barking, or else if barking had been introduced in the immediate linguistic context, in other words Evoked (Prince, 1981), or available via the episodic text-trace (Givón, 1995).

In Focus The most highly-activated referents are those that are not only activated (i.e. present in short term memory) but are also the current centre of attention, or **focus** of the utterance[8]; necessary for the appropriate use of zero

[8]Note the difference in usage of the term 'focus' here. Gundel *et al.* (1993) use 'focus' to refer to what the speaker assumes the addressee is currently attending to, what Hajičová (1991) would refer to as 'topic'. Hajičová uses the term 'focus' to refer to the information introduced in relation to the topic, that is, the new properties to be assigned, the modifications to be undergone, and/or the additional links to be created. Gundel *et al.* (1993) are aware of this potential confusion and discuss it in a useful footnote on page 279.

and unstressed pronominals. According to Gundel *et al.* (1993) the entities in focus at a given point in the discourse will be "that partially ordered subset of activated entities which are likely to be continued as topics of subsequent utterances" [page 279].

(4.13) *The dog next door is a real nuisance.* ***It*** *kept me awake all night.*

A comparison of Gundel *et al.* (1993)'s Givenness Hierarchy with Prince (1981)'s Familiarity Scale reveals that although some of the categories correspond, there are two important distinctions. Firstly, the Familiarity Scale does not distinguish between 'activated' and 'in focus': the term 'Evoked' covers both. Secondly, the relation between different levels on the hierarchy is different: the statuses in the Familiarity Scale are mutually exclusive, while the statuses in the Givenness Hierarchy involve entailment relations. However, the following more or less correspond with each other:

Prince's category	**Gundel *et al.*'s status**
Unused	'familiar' but not 'activated'
Containing Inferrable	'identifiable' but not 'familiar'
Brand New	'type identifiable' but not 'uniquely identifiable'

Whilst not all six of Gundel *et al.* (1993)'s statuses are required for all the languages they studied, it appears as if all pronominals and definite articles require the same statuses cross-linguistically. Further, forms which signal the most restrictive cognitive status ('in focus') are always those with the least phonetic content: unstressed pronouns, clitics and zero pronominals.

4.6.3 Accessibility and linguistic reduction: Lindblom revisited

From the work of Prince (1981), Ariel (1991) and Gundel *et al.* (1993) (and others cited therein) it is clear that there is a strong relation between the cognitive status of a referent (or a speaker's beliefs about the cognitive status of a referent for her listener) and the linguistic choices made by the speaker in referring to it. The more 'activated' or 'accessible' the referent, the less informative, more ambiguous, and more reduced the linguistic form used by the speaker to refer to it. For referents that are 'in focus' (Gundel *et al.*, 1993) the speaker has the option of using a maximally reduced expression, the zero pronoun (ellipsis). For less accessible referents, the speaker can select from a number of more elaborated expressions.

This relation between accessibility and reduction in size of referring expression would appear to reflect a similar principle of "Least Effort" to that proposed by Lindblom (1983a): the speaker only produces a long descriptive NP when it is required by her listener to successfully identify the referent. Indeed, both Prince (1981) and Gundel *et al.* (1993) appeal to Grice (1975)'s **maxim of Quantity** which states:

- Make your contribution as informative as required (for the current purposes of the exchange).
- Do not make your contribution more informative than is required.

Making a more informative contribution than is required would flout the maxim and result in an infelicitous utterance. (See Prince (1981)'s discussion of the Familiarity Scale and the effects of inappropriate use.)

But there is more to a speaker's choice of referring expression than simply adhering to Grice's maxim of not saying more than you have to. Lindblom's H & H theory suggests that speakers produce an utterance with as little articulatory effort as they can get away with. Ariel makes it clear, however, that the form of referring expression is not selected by the speaker simply in terms of what they can get away without saying. The degree of attenuation or reduction functions as information in itself[9], as a *marker* of how accessible the speaker thinks the referent ought to be for the addressee. Selecting the wrong form, producing an infelicitous expression, will send an incorrect signal to the listener, a signal to go and retrieve a representation that may not exist, or that exists in a different location than that implied by the speaker's choice of NP. It is not clear that Lindblom is arguing for a similar effect on the addressee when a speaker inappropriately hyper- or hypo-articulates, although the idea that poor articulation may have a signalling function has been suggested by others (Fowler and Housum, 1987). Indeed, there is an underlying assumption in Lindblom's H & H theory that 'speakers get it right' – that speakers are able to assess the needs of their listener from moment to moment and successfully adjust articulation accordingly. The notion of inappropriate location along the H & H continuum is therefore not addressed by Lindblom.

The question of how cooperative conversation interactants really are is discussed in more detail in Section 4.9. In the meantime I wish to focus on the relation

[9]See Ziv (1996)'s review of Lambrecht (1994) where he argues that Given $\neq$ redundant $\neq$ informationally irrelevant.

between information structure – the Given/New distinction – and possible effects on linguistic production and perception, with particular regard to the role of intonation and the articulatory correlates of NP repetition in discourse.

4.7 Information status as redundancy: effects on intelligibility and duration

It should be clear from the preceding discussion of the Given/New distinction that Givenness is one reflection of predictability or **redundancy**. A word that is Given, whether it be Given by previous mention, physical situation, or background knowledge, is likely to be more predictable (in context) than a word that is informationally New. Indeed, as we saw in Section 4.5, some early definitions of Givenness included the very term 'predictability' (Kuno, 1978). In this section, I highlight some general findings regarding the effects of redundancy on both production and perception, concentrating on measures of duration and intelligibility. After looking at redundancy as **predictability from sentential context** I will go on to consider research relating to redundancy defined in terms of **repeated mention** or (Textually-)Evoked Familiarity, i.e. Givenness (Prince, 1981).

The idea that the production and perception of any word in a meaningful sentence is a function of the speaker's and listener's knowledge of the information conveyed by the entire utterance is not a new one. In an experiment which explored the effects of 'degree of predictability from sentential context' on both duration and intelligibility, Lieberman (1963) found that tokens taken from well-known adages were shorter and less intelligible than tokens of the same word form taken from less predictable contexts. Thus, in (4.14), the word *nine* excerpted from (4.14a) was shorter and harder to recognise in isolation than the token of *nine* excerpted from utterance (4.14b).

(4.14) a. *A stitch in time saves* ***nine.***

b. *The number that you will hear is* ***nine.***

Two sets of test sentences were used: an idiomatic phrase in which the target word was highly predictable (redundant) was paired with a sentence in which the same target word was not predictable (non-redundant). Each word's redundancy was measured by asking readers to fill in the blank in the sentence left by the word when removed: the percentage of correct guesses constituted the index of

redundancy. The target words were then excised from the recorded productions of three speakers, masked by noise, and presented to a group of listeners to recognise.

Lieberman showed that the redundant tokens were identified less often than their non-redundant partner. In addition the production of the redundant tokens had clearly been attenuated: word duration and relative peak amplitude were reduced compared with the non-redundant tokens. Lieberman interpreted his results as indicating that "the speaker calls attention to the words that he thinks are non-redundant" (Lieberman, 1963, page 181). The converse of this is, of course, that in production a speaker may utter a word with less care when she knows that her listener can identify the word from its context. The parallel with Lindblom's H & H theory is obvious.

Although Lieberman (1963)'s findings have been much cited, they have also been criticised, primarily for a lack of data (the set of sentence pairs used was only eleven, from which just seven words were analysed) and for failing to control for aspects such as sentence position and phonetic environment. In a study designed, in part, to replicate Lieberman's work, Hunnicut (1985) compared the intelligibility of tokens extracted from two sets of sentence pairs which varied in redundancy value. One set of sentence pairs involved high redundancy adages matched with a corresponding set of grammatically similar sentences with low redundancy contexts, that "might be spoken" (Hunnicut, 1985, page 48). The other pair of sentences involved rather long, grammatically standardised sentences of the sort to be found in written texts, which were matched for grammatical structure, but where the redundancy was either high or low. (It should be noted that there is no indication of what the mean redundancy or standard deviation was for the various sentence sets). All sentences were in Swedish. There were 148 sentences in all, read by one male Swedish speaker in randomised order. Test words were edited out, overlaid with 'pink'[10] noise, and then played to ten subjects to identify.

Results showed that for the 19 text-type sentence pairs there was an intelligibility advantage for the words in lower redundancy contexts (mean intelligibility for high redundancy = 38.2%, mean intelligibility for low redundancy = 46.3%, difference significant at 2.2% level, paired comparison test). No such effect of redundancy on intelligibility was found for the 21 adage/spoken-type sentence pairs. Hunnicut suggests that the use of adages was, perhaps, a poor choice, in that their metaphorical nature, and usage in particular social situations, may result in their being far from redundant in the sense of "saying something everyone

[10]Unlike white noise, pink noise has certain speech-like characteristics.

knows already" (1985:53). She concludes that the intelligibility advantage found for words in lower redundancy contexts in text-type test sentences indicates that some information control is indeed undertaken by the speakers even for reading test sentences.

If, then, the grammatical and semantic information available from sentential context can affect the degree to which a word's production is attenuated, what might be the effect of *pragmatic* information, and specifically, knowledge that a word has been mentioned before? Does previous mention, or Givenness, increase redundancy and lead to more attenuated production?

4.7.1 The Repetition Effect

A production experiment which, again, was designed to replicate and extend Lieberman (1963)'s findings, was conducted by Shields and Balota (1991), in which the duration and peak amplitude was measured for target words produced in one of three sentence conditions. Target words could be:

- repetitions of a word mentioned earlier in the sentence
- associatively related to an earlier word (e.g. *shoes – socks*)
- unrelated to any earlier word

For the example in (4.15), the target is the fifth word in the sentence (*cat*) and the prime, which replaces the vacant slot, is one of the set comprising [*cat* (repeated), *dog* (associatively related), *son* (unrelated)].

(4.15) *Her ___ chases our cat under the table.*

Eighteen subjects read 18 sets of sentences, each of which had three prime-target pairs. The 54 sentences were written on separate pieces of card and presented in random order. Subjects were asked to read the sentence silently to themselves, and then convert the sentence from present- into past-tense form. To prevent them from *reading* the sentences directly, subjects were asked to place the card on which the sentence was written face down before producing their utterance. Although there was no real listener present, subjects were instructed "to produce each sentence as if they were relating the information to someone, using a normal speech rate and loudness level" (1991:50).

The results showed significant effects of repetition on duration and amplitude and a significant effect of relatedness on duration. Unrelated words were longer than associatively related words, which in turn were longer than repetitions, while repeated tokens had a significantly lower peak amplitude than either the associated or unrelated words (which did not differ from each other). Repetition of a word token clearly results in a reduction in duration and peak amplitude. Shields and Balota also conclude that "the mere presence of an associatively related word earlier in the sentence context is sufficient to modulate production durations of a target word" (1991:52). As in the case of word-form repetition, the results for associatively related words can be interpreted as an effect of predictability: the association between *cat* and *dog* is a relation of Givenness_k, a fact of general knowledge about the world to which a speaker would believe her listener had access.

It should be noted, here, that Shields and Balota (1991)'s results demonstrate an impact of repetition for repeated words with **different referents**: in (4.15) above, *her cat* and *our cat* refer to two different entities. Repetition effects elsewhere in the literature reveal conflicting findings with respect to referential identity. For example, Bard *et al.* (1989) found that repetitions, or self-corrections, which were **co-referential** and therefore added nothing New to the discourse, were significantly shorter and/or less intelligible than the original (first) mention; in contrast, if the repetition or self-correction introduced a New discourse entity or provided other information that was New to the discourse, then no shortening or loss of intelligibility was found.

Communicative context has also been demonstrated to influence the degree to which repeated words undergo durational shortening (Fowler, 1988). Fowler conducted a series of three experiments – the first using word lists, the second paragraphs of 'meaningful prose', and the third spontaneous conversation – in order to establish whether it was mere repetition that was important, or that it was indeed information redundancy that was responsible for durational shortening. She found no durational shortening for tokens repeated within word lists, a small but significant amount of shortening for tokens in paragraphs, and a large effect on duration for repetition in spontaneous conversation, demonstrating that repeated-word shortening is *not* a consequence of repetition alone, but depends upon a situation in which words occur in meaningful utterances.

In a related study, Fowler and Housum (1987) compared both the duration and intelligibility of first mentions with subsequent mentions of words taken from spo-

ken monologues. Subsequent – or Given – mentions were significantly shorter, and less intelligible when presented in isolation, than mentions which introduced the New token into the discourse. Fowler and Housum infer from their results that "talkers may attenuate their productions of words when they can do so without sacrificing communicative efficacy" (1987:489). In other words, acoustically less informative versions of words may be uttered when there is good supportive context making the word more probable than in a low probability or non-redundant context. Because words which are Given receive greater contextual support they can be reduced without loss of understanding by the listener.

In two additional experiments, Fowler and Housum (1987) show how listeners were able to identify Given and New words as such, and use this knowledge of a word's being Given rather like using an anaphor. Words were presented in isolation and subjects asked to identify each word as either 'New' or 'Old' (=Given). Although subjects found the task difficult and made a number of errors, almost all of them could do it. The average success rate of 60% was significantly greater than chance. Fowler and Housum then asked whether the knowledge that a word was Given could actually promote comprehension by informing the listener that the word refers back to something said previously. They hypothesised that a reduced token of a spoken word serves as a better reminder of words in a sentence containing a non-reduced token of the same word form, than does the non-reduced token itself. Measures of Reaction Time to prime-target word pairs in a prime-probe experiment showed that responses to reduced primes were faster than to non-reduced primes: the token's reduction appears to facilitate recall of the word's prior context.

In summary, the repetition of a co-referential word in context leads both to reduction in duration and peak amplitude, and to a loss of intelligibility when the word token is excerpted from context, masked with noise, and played back to a group of listeners to identify. Assuming that repeated mention of words within a discourse is an example of Givenness (Prince (1981) refers to repetition of this sort as Textually-Evoked Familiarity), there is clearly a relation between Givenness, duration and intelligibility: words which refer to Given information are shorter and less intelligible than words which refer to information which is New. This relation has been interpreted as reflecting an effect of redundancy on production: the more informationally redundant a word token is, the more attenuated is its production. Thus far, the empirical findings appear to support Lindblom's general contention that speakers reduce articulatory effort according to their beliefs about the contextual information available to the listener to help decode the sig-

nal. In the following section I turn to the influence of prosody in distinguishing New from Given, and focus on the relation between Given/New and the location of pitch prominence, or accent, since it has been argued (Hawkins and Warren, 1994) that the Given/New repetition effect on duration and intelligibility is, in fact, an artefact of a difference in accent status, rather than information structure *per se.*

4.8 The Prosodic marking of Given/New

It was Halliday (1967) who first appealed to the Prague School approach to information structure, and specifically their Given/New distinction, to account for intonational patterns in British English. Halliday was interested in the organisation of information in spoken English, and specifically its phonological realisation. He proposed that the contents of a clause – the basic unit in his grammatical system – are organised into one or more **information units** which are realised phonologically by intonation. Information units are associated with **tone groups**, with each tone group containing one, and only one, **tonic syllable**, the syllable with the maximal pitch movement. This tonic syllable functions to focus the New information in the tone group. Halliday suggested that, characteristically, speakers order Given information before New information. Therefore, in the *unmarked* case, the tonic syllable will focus the last lexical item in the tone group, which will generally be the head-word of the constituent containing New information. According to Halliday's classification, every tone group must include a chunk of New information, which will be phonologically marked by the tonic pitch movement. The speaker has the option of including one or more chunks of Given information, which, by definition, will not be phonologically marked by pitch prominence.

Central to Halliday's account is the role of the speaker in deciding what to treat as New and what as Given: the assignment of Given/New status to information is determined by the speaker, not by the text. Thus, the status of information is *not* dictated by whether or not an entity has been referred to already within the discourse.[11] Halliday observes:

> *"These are options on the part of the speaker, not determined by the textual or situational environment; what is New is in the last resort what the speaker chooses to present as New, and predictions from the*

[11] Chafe (1976) shares this view. He states that "givenness is a status decided on by the speaker" (Chafe, 1976, page 32), that it is "fundamentally a matter of the speaker's belief that the item is in the addressee's consciousness" (*ibid.*).

discourse have only a high probability of being fulfilled."
(Halliday, 1967, page 211)

In theory, then, a listener ought to be able to determine how an utterance is chunked into tone groups by attending to the intonation contour: information within a focused domain is New if the tonic is present, and Given if it is absent. In practice, however, there are problems, as Brown (1983) observes. Firstly, the intonational criteria are inadequate for the identification of information units in speech (Brown *et al.*, 1980). Neither tonics nor tone groups can be consistently identified (*ibid.*:29). Secondly, there are problems with insisting that a tone group be associated with *only one* tonic, or New information marker (see Brown, 1983, page 68, for references). Nevertheless, the assumption that intonational prominence has, as one of its functions, the role of focusing the listener's attention to particular parts of a speaker's message, is one that is now generally held by most researchers in the area (Cutler, 1976; Bolinger, 1978; Bock and Mazzella, 1983; Fowler and Housum, 1987; Nooteboom and Kruyt, 1987; Terken and Nooteboom, 1987; Walker, 1992; Hawkins and Warren, 1994; Vallduví and Zacharski, 1994; Fisher and Tokura, 1995).

As has been discussed above, in Section 4.5, Halliday's categories of Given and New have been extended in the literature to include all aspects of what a hearer might be predicted to 'know', whether from the discourse context, or from sources external to the current context of utterance. As a result, the status 'Given' has been assigned to forms which would inevitably attract phonological prominence, claims Brown (1983). The relation between pitch prominence and Givenness depends, therefore, on a researcher's operational definition of what it means for something to be assigned the status Given. For Halliday, Givenness is defined in terms of (lack of) pitch prominence, and the two are therefore inseparable. More recently, however, Givenness has tended to be treated as a property of the text; in particular, repeated mention of an item in a small (usually two-sentence) 'discourse' has been taken to refer to a Given entity, while the first, introductory mention has been assumed to be New. Adopting an independent definition of Givenness like this has allowed researchers to examine more closely the relation between Givenness – or, more accurately **repetition** – and pitch prominence.

4.8.1 Given/New as +/− Accent

Rather like the terms 'Topic' and 'Focus', *Accent* has a number of different interpretations in the literature, depending on the intonational 'school' from which the

research originates (see Cutler and Ladd (1983) for a discussion). In this thesis, I will adopt a similar usage to that of Hawkins and Warren (1994) whereby the term **Accent** will be used to mean *any stressed syllable with* ***pitch prominence***, or, more specifically, a ***prominence-lending pitch movement***, and **Nuclear accent** to refer to that accented syllable (usually the last in a tone group) carrying the ***major*** *pitch movement* within an intonational phrase. Accented words will be written in SMALL CAPITALS.

It is generally accepted that the presence or absence of accents – or prominence-lending pitch movements – helps hearers to distinguish between the important, relevant speech fragments, and those parts of the utterance which are less relevant. In a phoneme-monitoring task, Cutler (1976) found, for example, that listeners were faster in detecting initial phonemes in accented words than in unaccented words, suggesting that the presence of an accent speeds up the use of acoustic-phonetic information in the recognition of words.

Various studies – on both production and perception – have demonstrated a strong relation between accentuation and information structure: informationally New items are associated with pitch accents, while reference to informationally Given entities tends to be deaccented (Bock and Mazzella, 1983; Brown, 1983; Terken, 1985; Nooteboom and Kruyt, 1987; Terken and Nooteboom, 1987).

For example, Terken (1985) found that deaccenting is strongly associated with previous mention: referring expressions that had been mentioned in the previous utterance were more likely to be deaccented than referring expressions which were introduced for the first time. Terken also showed that comprehension is facilitated by the appropriate distribution of accents: words containing Given information were verified faster when they remained unaccented than when they were accented, while words containing New information were verified faster when they were accented. The verification paradigm involved listeners watching a picture (a configuration of several alphabetic characters) on a computer screen; aspects of the picture were altered by moving a character to a new position; each change in configuration was followed by a spoken description of the alteration, such as "*the p is on the right of the k*"; subjects had to decide as quickly as possible whether the description was True or False. It has been argued from the results of such verification tasks (Terken, 1985; Terken and Nooteboom, 1987) that listeners do not simply give more attention to accented words, but that they process them differently from unaccented words: deaccentuation has a communicative function in its own right, serving to mark an expression as non-focal and requiring different

treatment. Leaving an expression unaccented at the right time is as important as accenting an expression at the right time (Terken, 1985, page 123).

In a perceptual experiment exploring the relations between accent, focus distribution and Given/New, Nooteboom and Kruyt (1987) varied accent patterns systematically by manipulating their synthetic speech contours. Subjects used a rating scale to judge the 'acceptability' of different contours in combination with different leading sentences which determined the distribution of Given and New information. Accented target words were rated as most acceptable when referring to New information, and unaccented target words were most acceptable when referring to Given information.

More recently, Eefting (1991) investigated the effects of accentedness and information value on **duration**, in a production experiment which made use of the fact that only one word in a constituent can carry the nuclear accent, despite the fact that several words within the constituent might contain New information. She used 16 target words altogether; 8 of these were monosyllables with a CVC structure, and 8 were tri-syllabic words, with a Cə(C)-C(C)V-CəC structure. (The vowel was always the last segment in the lexically stressed syllables of the polysyllabic words.) For each set of characteristically 'short' and 'long' words, half contained the long vowel [a:] and half the long vowel [e:], with the [a:] and [e:] words paired and matched as closely as possible for phonetic context, e.g., *kaas* (meaning *cheese*) was matched with *kees* (the proper name *Kees*); *gekakel* (meaning *cackling*) was matched with *gebedel* (meaning *begging*) *etc.* The surrounding consonants were chosen to be easily detectable on an oscillogram, in order to facilitate measurement.

The target words were used in the construction of 3 sets of stimuli. In the first set, sentence pairs (matched for length) varied the accent status of the target word while the information value was held constant, in this case always New. (See examples (4.16a) and (4.16b).)

In the second set, target words were embedded in the last sentence of a two sentence fragment. The target word was always unaccented. The immediately preceding sentence was then manipulated so that in one condition the target word contained New information (such as in (4.17a)), while for the second condition it contained Given Information (as in (4.17b)).

(4.16) a. *Gisteren hebben* KEES *ten* KATE *en* MARIE *van der* BILT *een* PRIJS *in de* LOTERIJ *gewonnen.*
(*Yesterday* KEES *ten* KATE *and* MARY *van der* BILT *won a* PRIZE *in the* LOTTERY.)

b. *Gisteren heeft Kees ten* KATE *de* HOOFDPRIJS *in de* STAATSLOTERIJ *gewonnen.*
(*Yesterday Kees ten* KATE *won the* FIRST PRIZE *in the* STATE LOTTERY.)
target word = "*Kees*"

(4.17) a. i. *Wat zei je over Kees ten Kate?*
(*What did you say about Kees ten Kate?*)

ii. *Gisteren heeft Kees ten Kate een* PRIJS *in de* LOTERIJ *gewonnen.*
(*Yesterday Kees ten Kate won a* PRIZE *in the* LOTTERY.)

b. i. *Wie heeft er gisteren een prijs in de loterij gewonnen?*
(*Who won a prize in the lottery yesterday?*)

ii. *Gisteren heeft Kees ten* KATE *een prijs in de loterij gewonnen.*
(*Yesterday Kees ten* KATE *won a prize in the lottery.*)
target word = "*Kees*"

The third set of stimuli combined the effects of information value and accent status, using two conditions: [+accent, New] and [−accent, Given]. Eefting excluded the conditions [−accent, New] and [+accent, Given] on the grounds that these combinations were less acceptable (Nooteboom and Kruyt, 1987), and also increased processing time (Terken, 1985), which might affect duration values. The two sentence pairs for "*Kees*" in the combined condition can be seen in (4.18a) and (4.18b).

(4.18) a. i. *Wat zei je over Kees?*
(*What did you say about Kees?*)

ii. *Gisteren heeft Kees een* PRIJS *in de* LOTERIJ *gewonnen.*
(*Yesterday Kees won a* PRIZE *in the* LOTTERY.)

b. i. *Wie heeft er gisteren een prijs in de loterij gewonnen?*
(*Who won a prize in the lottery yesterday?*)

ii. *Gisteren heeft* KEES *een prijs in de loterij gewonnen.*
(*Yesterday* KEES *won a prize in the lottery.*)

One professional reader read the set of 96 stimuli aloud five times resulting in 480 target word productions, which were then measured. Eefting found that while accentuation had a significant effect on duration – accented words were roughly 25% longer than unaccented words – there was *no effect of information value*: Given words were just as long as New words.[12]

Measurements of the syllable and segment durations within each word revealed that *all* segments and syllables contribute to the durational changes that arise through accentuation, although not to the same degree. In the monosyllabic words, vowel duration appeared to contribute less to the durational difference, than the initial and final consonants. In the polysyllabic words, the segments forming the lexically stressed syllable appeared to show the greatest change in duration (although there was no statistical support for this observation).

Eefting contrasts her findings with those of Hunnicut (1985) and Fowler and Housum (1987) (amongst others) who *did* find durational consequences of information status, and argues that they failed to vary the factors 'accent' and 'information value' independently; it is not clear therefore whether the durational effects found by Fowler and Housum (1987) *etc.* were caused by accentuation rather than information status *per se*.

Although Eefting (1991) is cautious in her conclusions, suggesting, for example, that her read speech tokens may differ from natural conversation, and that other aspects of production such as intensity and pitch movement may play a role, she fails to acknowledge the biggest problem with her study, which is the incompleteness of the stimulus set. In order to vary accent and Givenness *independently*, as she proposes, it is necessary to explore what happens when accent is held constant while information status is varied, and vice versa; in other words, a '2x2' factorial design is called for, with the values +/−Accent and +/−Given (or −/+New). Eefting manages to fill only three of the four cells: there are no cases of [+accent, +Given]. This is evident from her own table (Eefting, 1991, page 416), reproduced here as Table 4.1, which shows that she could examine Given/New effects only within unaccented tokens.

Though she finds no effect of Givenness on duration for unaccented tokens (using stimuli from set II), she cannot tell us whether there is durational shortening for Given words which are accented.

Though Eefting argues that accented Given words are less acceptable (Noote-

[12]No effects were found for vowel or word length (once number of syllables was controlled for by looking at difference measures).

	+Accent	−Accent
+New	**set I**	**set I**
(−Given)	*set III*	SET II
−New	—	SET II
(+Given)		*set III*

Table 4.1: Eefting's experimental design for investigating the relation between accent and information structure (adapted from Eefting (1991))

boom and Kruyt, 1987) and slower to verify (Terken, 1985) than unaccented, she cannot argue that accented Given words do not occur. Terken (1985), for example, found that the percentage of accented expressions remained quite high even for Given referents. It was only as the number of consecutive utterances in which the referent appeared grew that speakers started to use unaccented referring expressions.

Horne (1991) has similarly observed that although it is more or less the rule that New information is accented in neutral intonation, "it is not the case that Given information is deaccented" (page 1279). She argues that the accenting of Given information can be explained in terms of general metrical well-formedness conditions on prosodic constituents. Comparing the metrical restrictions on the structure of words in English, Horne draws a parallel with the patterning of accents on prosodic levels higher than a word. She concludes that the deaccenting of Given information only occurs when the Given material occurs in post-focal or post-nuclear position in a phrase, in other words, when it comes *after* the last New (and consequently accented) word in an intermediate phrase. Thus in (4.19a) the Given material (*bastards*) referring to *The children* in (4.19) is deaccented, while in (4.19b), it is likely to be accented, because it precedes New, accented information (SCOLDED).

(4.19) *The children didn't want to go to bed so*

a. *John* SCOLDED *the bastards.*

b. *the* BASTARDS *were* SCOLDED.

Ladd (1996) offers a similar example (4.20), where deaccenting distinguishes between an 'epithet' and 'literal' interpretation of *butcher*. In (4.20a) deaccenting makes it possible to treat *butcher* as an epithetical reference to the surgeon who performed the operation. When *butcher* is accented the reply becomes an inco-

herent rant about a literal butcher who sells meat. In other words, by accenting the post-nuclear item a New referent is introduced to the discourse.

(4.20) *Everything OK after your operation?*

a. *Don't talk to me about it! I'd like to* STRANGLE *the butcher!*

b. *Don't talk to me about it! I'd like to strangle the* BUTCHER*!*

Example 4.19 (originally from Berman and Szamosi (1972)) is discussed by Nooteboom and Kruyt (1987) who, like Terken (1985) and Horne (1991), find Given words can be accented and still sound acceptable. While accented constituents are generally judged as most acceptable when referring to New information, under certain circumstances, "a constituent can acceptably be focused [=accented] also when it refers to Given information" (page 1520). Their data demonstrates an asymmetry: lack of accent can rarely be acceptably associated with New information[13] but accenting can often (though not always) be associated with Given information. This is not surprising, as the function of intonation is not exclusively a matter of demarcating New from Given information. Intonation is a focusing device which can highlight, for example, the theme or topic of a sentence, as well as what is informationally New or Given. Furthermore, rules governing accent assignment mean that constituent structure and word order will also play a role in determining whether a particular word is accented. For example, default accent placement can result in ambiguous focus domains, as in (4.21) where either the whole constituent, or only "*our* FACULTY" is in focus.

(4.21) *The dean of our* FACULTY

In such cases, the lack of accent on *dean* is found to be equally acceptable whether *dean* refers to New or Given information (Nooteboom and Kruyt, 1987).

More recently, Walker (1992) has shown not only that Given information may be accented, but that the nature of a word's Givenness (Given as Known versus Given as Salient) may influence the choice of intonational contour. Walker explored the different intonational realisations of utterances which consist of wholly Given information, which she refers to as **Informationally Redundant Utterances**, or IRUs. She distinguishes four logically different types of IRU, according to how they relate to their antecedent: repetition, paraphrase, entailment, and non-logical inference. An independent intonational description, based on that

[13] but see the discussion of example (4.21) below

proposed by Pierrehumbert (1980), revealed that while IRU type predicted the scope of focus (inferrable IRUs are more likely to be realised with an item in narrow focus than IRUs classified as repetitions or paraphrases), it did not predict contour type. The best predictor of contour type was found, instead, to be *saliency*. IRUs in which the antecedent was no longer salient, that is, when the antecedent had been displaced by an intervening change in topic, were never produced with sustained tones. Walker was unable, from her data, to predict when sustained tones should occur, only when they should not.

To summarise this section, then, there is clearly a relation between accent and Given/New, but the relation is asymmetric. While New words are almost always associated with prominence lending pitch movements, the accent status of Given words is more variable. When Given words are unaccented, they are verified faster and perceived, generally, to be more acceptable than when they are accented. However, there are a number of circumstances in which accenting Given information does not violate acceptability. In a production experiment exploring the relation between accent, Givenness and duration, Eefting argues that it is accent rather than information status that effects changes in duration. Poor experimental design, however, leaves open the question of whether accented words referring to Given information are shorter than accented words referring to New information.

4.9 Is language always cooperative?

It is evident from the preceding sections that much of the work on discourse, and particularly dialogue, assumes a high level of cooperation between speaker and listener. Dialogue participants are presented as collaborators who establish common ground together by participating in both presentation and acceptance phases of a contribution until they arrive at a mutually held belief (Clark and Schaefer, 1987a).

Similarly, speakers produce referring expressions that appropriately mark the accessibility of the referent for their listener (Ariel, 1990; Gundel *et al.*, 1993), which thereby saves processing time (Ariel, 1991; Givón, 1995): if accessibility guides referent retrieval, then the information relating to the referent will be (re-)activated more quickly. Terken and Nooteboom (1987) suggest that deaccenting plays a similar role: the attenuation associated with unaccented words (shorter duration, lower peak amplitude) functions to signal to listeners that they should process

such words differently. That is, attenuation in articulation can be seen as part of a continuum of reduction, where the degree of reduction serves as an accessibility marker, indicating where in the mental representation of discourse to search for the appropriate referent.

Viewed in this light, the adjustment of articulatory clarity up or down the Hyper/-Hypo-speech continuum (Lindblom, 1990a) has benefits for both speaker and listener. The hypo-articulation associated with predictable information is less effortful for the speaker, but also valuable to the listener, the attenuation in the signal indicating where in the discourse representation to seek the appropriate referent.

The picture presented here is one of a caring, considerate speaker who appears to empathise with her listener and cooperate with him to the best of her abilities. But just how accurate a depiction of speaker behaviour is this?

There is some evidence in the literature to suggest that not all speaker behaviour is quite so self-less and listener-oriented. As Anderson *et al.* (1997) observe, it would require "extraordinary powers" on the part of the speaker to develop and update the detailed model she would require to assess her listener's current knowledge requirements. Perhaps, rather than attempting to model her listener's current discourse state, the speaker makes the assumption that *her own* discourse model will be sufficiently similar to her listener's to count as "the same"? In this case she will base her beliefs about the Givenness and accessibility of a referent via her own model rather than one she has constructed for her listener.

In the majority of situations this strategy may well prove adequate. However, sometimes mistakes will occur. In Section 4.7.1 above it was shown that repeated mentions of entities within a discourse were shorter and less intelligible than introductory mentions. When Bard and Anderson (1994) looked at repetition in adult speech to children, they found the same loss of intelligibility for repeated mentions, notwithstanding the reason for the repetition: the inattentiveness of their listener. Adults repeated themselves precisely because the child had failed to attend to the earlier utterance. Consequently, the poorly articulated repetition was presumably New for the listening child, although it was Given for the adult speaker.

A similar un-Gricean effect was found for speech in recorded dictations. Bard *et al.* (1989) showed that subjects using a dictation machine produced degraded second tokens despite the fact that they had just actively erased the first token in reformulating the memo. The attenuated token was therefore the first mention

to be heard by the audio-typist.

In both these examples the speakers' productions were related to the status of Given/New in their own model of discourse, rather than their listener's, making them rather less cooperative than the ideal speaker presented by Lindblom and others.

Chapter 6 describes a series of intelligibility experiments which were designed to reveal how sensitive speakers really are to the informational needs of their listener.

4.10 Summary

Lindblom (1990a) has argued that the economy with which a speaker articulates a token is influenced by her beliefs about the information requirements of her listener. This chapter explored the nature of the signal-independent information available to the listener to help him decode the acoustic input: the contextual information which Lindblom suggests affects the degree of hyper/hypo-articulation.

In particular the chapter focused on the distinction between 'Given' and 'New', and detailed some of the linguistic means by which this distinction is conveyed. It was shown that repeated reference of an entity results in shorter, more degraded tokens compared with the first, introductory token. Givenness, then, is associated with attenuated production. A relation was also demonstrated between Givenness and sentence Accent: whilst reference to New entities involves a pitch accent, reference to what is Given is frequently (though not always) deaccented.

Chapter 5

Materials and Methodology

5.1 Introduction

In this chapter I describe the source of speech data used to test Lindblom's claims about articulatory effort. Lindblom's H & H theory argues that articulatory effort is varied according to the informational requirements of the listener. To test such a hypothesis requires speech data that involves both speakers and listeners in real communicative situations. With such material we can be sure that there is a genuine communicative purpose – to exchange information – and that the listener's informational needs will vary according to what has occurred earlier in the discourse. However, as we observed in Section 2.4.2.2, there is a problem with naturally occurring unscripted speech: critical aspects of both the linguistic and extralinguistic context may be either unknown or uncontrolled. Further, there is no guarantee that the linguistic phenomena of interest will appear with sufficient frequency to support an objective, quantifiable analysis. While prepared materials may lack spontaneity, they have the advantage of being designed to elicit specific examples of linguistic behaviour in controlled conditions, ensuring that the particular research needs are met.

What is required, then, is a corpus of unscripted dialogues that were elicited in such a way as to boost the likelihood of occurrence of particular linguistic phenomena, while also controlling some of the effects of context. The **HCRC Map Task Corpus** (Anderson *et al.*, 1991) fulfils exactly these requirements: while the dialogues are themselves entirely unscripted, or 'spontaneous', the corpus as a whole essentially comprises a large, carefully controlled elicitation exercise.

In the following section I provide a general description of the Corpus, focusing on the variables that are pertinent to the research described in the ensuing chapters.

I follow this with a detailed account of the segmentation procedure I followed when extracting the speech material required both for presentation in the intelligibility experiments described in Chapter 6, and for the duration analyses detailed in Chapters 8 and 9.

5.2 The HCRC Map Task Corpus

5.2.1 Task description

The Map Task (Brown *et al.*, 1984) is a cooperative task involving two participants. The two speakers sit opposite one another and each has a map which the other cannot see. One speaker – designated the **Instruction Giver** – has a route marked on her map; the other speaker – the **Instruction Follower** – has no route. The speakers are told that their goal is to reproduce the Instruction Giver's route on the Instruction Follower's map. The maps are not identical and the speakers are told this explicitly at the beginning of their first session: "The maps were drawn by different explorers". It is, however, up to the two speakers to discover how the maps differ. There was no time constraint imposed on completing the task.

5.2.2 Map design

All maps consist of landmarks – or **features** – portrayed as line drawings and labelled with their intended name. Figure 5.1 depicts a Giver/Follower pair of maps. The differences in the maps arise from systematic manipulation of a design variable referred to as **sharedness**: the extent to which the features contrast or are shared between pairs of maps. Features were deemed common – or **shared** – if the identical form and label appeared in the identical location on both the Giver's and Follower's map. Features which were not common differed in one of three ways:

Absent/Present features were found on one map but not the other;

Name Change features were identical in form and location but had different labels on the two maps[1];

[1]Name-change landmarks appear on only half of the map pair sets. Maps belonging to Quads 1 and 2 have name change features; maps from Quads 3 and 4 do not.

2:1 features appeared twice on the Giver's map, once in a position close to the route and once more distant, while the Follower had only the distant *irrelevant* one.

In the example maps in Figure 5.1, shared landmarks include the *extinct volcano*, *rope bridge*, and *Saxon barn*; unshared landmarks include the *tribal settlement*, *machete* and *pelicans*. The 2:1 feature is *golden beach*: the *golden beach* which forms the route pivot-point in the top left corner is on the Giver's map only, while both Giver and Follower have an irrelevant *golden beach* on the other side of the map, above and to the right of *secret valley*.

All map routes begin with a starting cross labelled "START", marked on both maps, and end with a finishing cross marked only on the Giver's map. Both start and end points are adjacent to a common feature but landmarks between these points alternate in sharedness. Maps contained an average of 12 landmarks each, of which at least three were unshared. It is this problem of unshared knowledge that presents the major difficulty in replicating the route. No one participant has access to all the information required and consequently both partners need to exchange information about what they can or cannot see.

5.2.2.1 Landmark names

Since the only constraint on the range of map landmarks is the ease with which the feature can be represented graphically (that is, choice is restricted only by the ingenuity of the artist), it was possible to incorporate landmark names of specific phonological interest. Four phonological reduction processes were selected, and landmark names designed to provide the appropriate phonological conditioning contexts. The four reduction processes were:

- **/t/-deletion**
  e.g. *extinct volcano* may be pronounced as [ɛkstɪŋk # vɒlkeno][2];
- **/d/-deletion**
  e.g. *submerged rocks* may be pronounced as [sʌbmɛrʤ# rɒks];
- **glottalisation**
  e.g. *secret valley* may be pronounced as [sikrɪʔ # valɪ];
- **nasal assimilation**
  e.g. *Saxon barn* may be pronounced as [saksəm # baɾn];

[2]Transcription is in Standard Scottish English to reflect the accent of the Map Task subjects

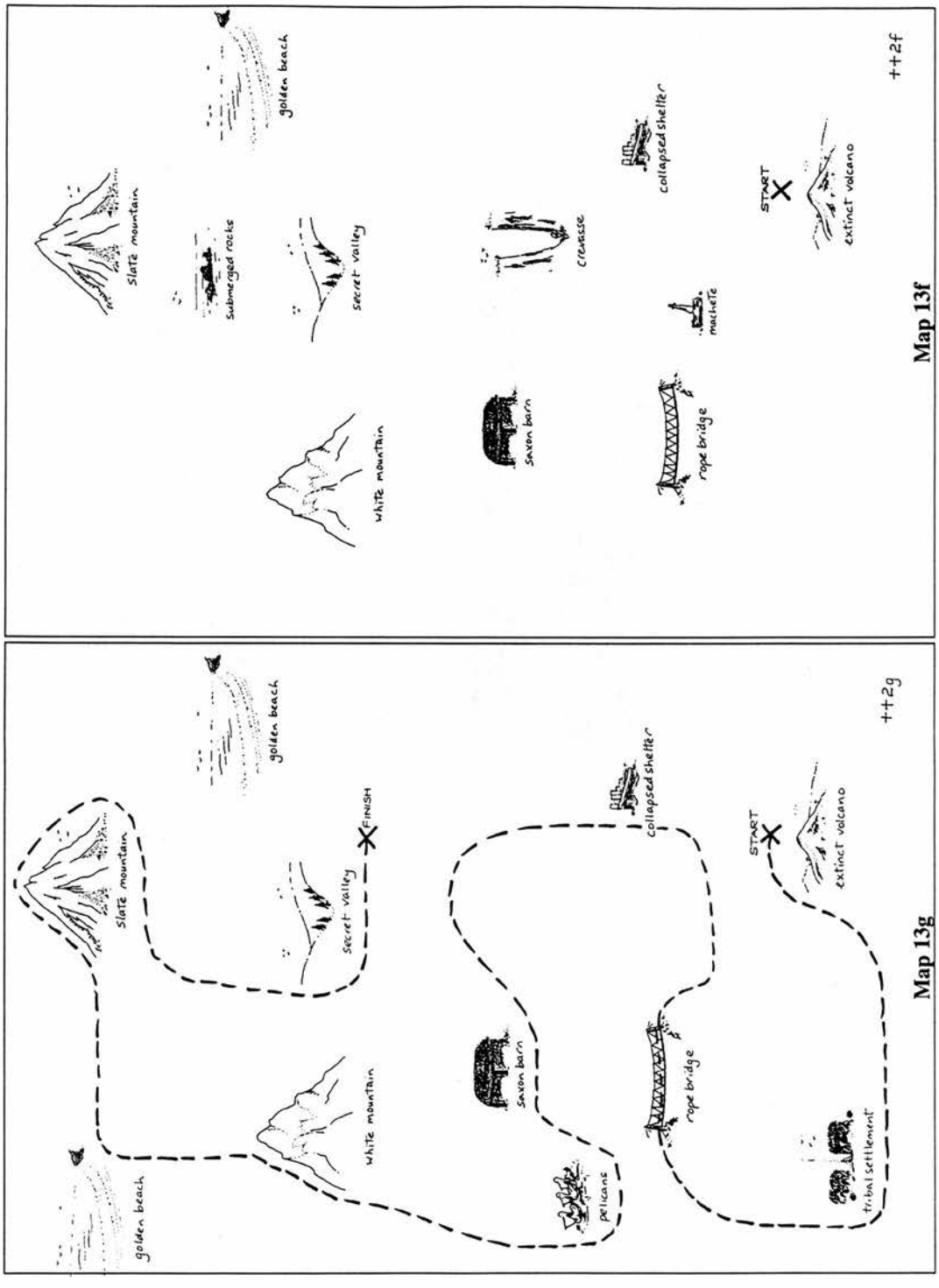


Figure 5.1: Example map pair from the HCRC Map Task Corpus
These maps were used in dialogues q4ec1, q4ec7, q8ec1, q8ec7, q4nc1, q4nc7, q8nc1 & q8nc7.

Landmark names also provided examples of polysyllabic words which differed in the metrical structure of the first two syllables: polysyllables could be Strong-Weak (henceforth **SW**) or Weak-Strong (**WS**), where 'Strong' refers to syllables with full (i.e. non-reduced) vowels which carry primary or secondary lexical stress, while 'Weak' refers to lexically unstressed syllables with reduced vowels ([ə] or [ɪ]). SW polysyllables included landmark names like *pelicans*, *elephants* and *telephone box*, while WS polysyllables included *machete*, *savannah* and *collapsed shelter*.

5.2.3 Corpus design

In addition to the design variables relating to the maps themselves, two other variables were incorporated in the design of the corpus overall.

Subjects are necessarily paired for the task, and since the pairing is under the experimenter's control it was possible to vary systematically the **familiarity** of the participants, by asking subjects to attend with a friend. Each pair of familiar subjects was run in coordination with another pair who were unknown to either member of the first pair. Two pairs formed a quadruple, or **Quad**, of subjects who used among them a different set of four map pairs, with maps being assigned to quads by Latin Square. Each subject participated in four dialogues, twice as Instruction Giver, and twice as Instruction Follower, once in each case with a familiar partner, and once with an unfamiliar partner. As Instruction Giver they gave directions on the same map, but when following they used different maps each time. Half of the subjects gave instructions to a familiar partner first, the others to an unfamiliar partner first.

The option of placing a small barrier between Map Task participants to prevent them form seeing each other's faces allowed us to control the availability of the visual channel for communication. Half of the subjects who took part in the task were able to make **eye-contact** with their partner, while the other half had **no eye-contact**. In both conditions participants were instructed **not** to use non-verbal gestures during the conversation.

The full design is presented in Appendix A. For further details see Anderson *et al.* (1991) or McAllister *et al.* (1990).

5.2.4 Subjects

The sixty-four subjects who participated were all undergraduates at the University of Glasgow. All but three of the subjects were Scottish, with 56 of them having been born or brought-up within a thirty mile radius of Glasgow. Two subjects were English, and one was North American. Half the subjects were male, half were female, and their mean age was 20 years old (age range was 17 to 30). All subjects were without any known speech or hearing impairment.

Subjects appeared to accommodate easily to the task and experimental setting, producing unselfconscious and relatively fluent speech.

5.2.5 Procedure

Subjects sat three or four feet apart, facing each other across a desk, with their maps placed on sloping boards, to prevent each subject seeing the other's map. Quads of subjects were randomly assigned to one of the two eye-contact conditions.

After they had completed their map dialogues, subjects were asked to read a word list containing all the feature names from the set of maps they had encountered. Feature names appeared twice in random order, and subjects were asked to read the list slowly and carefully, aiming for a between word interval of approximately one second. These list readings provided **citation forms** against which the unscripted dialogue forms could be compared.

Materials were recorded on Digital Audio Tape (sony DTC1000ES) using one Shure SM10A head-mounted microphone and one DAT channel per speaker. Split-screen video recordings were also made for one quarter of the corpus (Quads 3, 4, 7 and 8 in the eye-contact condition), which captured the face of the Instruction Giver and an angled view of the face of the Instruction Follower, along with his/her upper body and their map.

5.2.6 Transcription

All dialogues were transcribed verbatim in standard orthography, including (where possible) filled pauses, false starts, hesitations, repetitions and interruptions. These transcriptions were then checked several times against the original DAT recordings.

5.2.6.1 Dialogue excerpt

The following is an excerpt taken from the start of the first dialogue in Quad 4, between an unfamiliar pair of speakers, in the 'with eye-contact' condition. The Instruction Giver is speaker A, and the Follower speaker B, and they are using the pair of maps in Figure 5.1. The transcription notation for false starts, filled pauses *etc.* has been removed for ease of reading.

```
A: Start at the extinct volcano, and go down round the tribal
   settlement.  And then
B: Whereabouts is the tribal settlement?
A: It's at the bottom.  It's to the left of the extinct volcano.
B: Right.  How far?
A: Ehm, at the opposite side.
B: To the opposite side.  Is it underneath the rope bridge or
   to the lef
A: It's underneath the rope bridge.  And then from the tribal
   settlement go straight up towards the rope bridge and over
   the rope bridge.  Then down three steps and along to above
   the volcano.
B: Eh, d ... Is down three steps below or above the machete?
A: Ah. The machete's not on my map.
B: Oh.
```

5.2.7 Corpus statistics

The HCRC Map Task Corpus consists of roughly 16 hours of speech. The quantity of material involved is shown in Table 5.1. While the number of tokens is high, the task can be seen to have constrained the choice of word type (or form) to a pool of just under two thousand; in other words, there are multiple tokens of a small set of word forms.

5.3 Speech segmentation criteria

This next section details the segmentation criteria adopted for extracting the speech material from its source. Essentially, the problems of segmentation and duration measurement are the same (Peterson and Lehiste, 1960, page 694). The

	Total	With Eye-Contact	No Eye-Contact
No. of dialogues	128	64	64
No. of word types	1,939	1,489	1,469
No. of word tokens	146,855	66,729	80,126
Ave. word tokens per dialogue	1,147	1,043	1,252

Table 5.1: HCRC Map Task Corpus statistics

measurement of duration for any kind of speech segment or unit – be it sub-phonemic segments, whole syllables, or even words – is hampered by the fact that the 'beads-on-a-string' model of speech fails to capture the effects of coarticulation and the general overlapping nature of speech sounds.

However, in spite of the fact that phonemic segments cannot be said to occupy discrete, non-overlapping stretches of the speech waveform, it is nevertheless possible to produce transcriptions of utterances which are, within certain limits, time-aligned with the acoustic signal. This is because certain phonemic segments, such as the strident voiceless fricatives /s/ and /ʃ/ in English, have acoustic manifestations which are relatively stable and, in most phonetic environments, are easily identifiable from a spectrographic or similar display. There are, of course, segments or sequences of segments, such as vowels and sonorant consonants, which present genuine segmentation problems. On such occasions it may be necessary to locate the segment boundary more or less arbitrarily. But it should be remembered that segmentation can still be consistent, despite the application of arbitrary segmentation rules. Indeed, it is important to specify such rules explicitly, in order to ensure the comparison of like with like.

Word recognition experiments, such as the intelligibility studies undertaken in this thesis, necessarily require whole words to be excised from the speech waveform and presented to groups of subjects for a recognition response. This excision process is made more or less difficult depending on the degree of control the experimenter has over the phonetic context in which the word of interest occurs. In phonetic experiments involving the recording and analysis of individual words or nonsense CVC syllables, it has been traditional to use carrier phrases, such as *"Say X instead"*, or *"Say Y again"* (Peterson and Lehiste, 1960; Harris and Umeda, 1974) in order to control the phonetic environment. As well as avoiding pre-pausal lengthening effects by placing the word of interest in a non phrase-final position, the surrounding context can be selected to ensure an easy segmentation. The end point of a word that terminates with a plosive, nasal, or fricative can

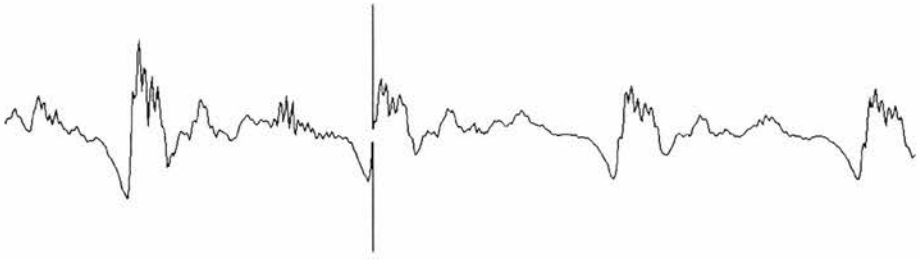


Figure 5.2: Time/amplitude waveform illustrating a segmentation boundary drawn at a zero crossing located on the 'up' stroke of a complex wave.

easily be located in the context of a following vowel, for example (hence the use of carriers such as *instead* and *again* above). Open CV syllables can be placed in carriers such as *"Say X several times"* where the onset of frication for the /s/ of *several* clearly defines the boundary that marks the end point of the CV syllable.

However, experiments involving the duration measurement of segments in spontaneous discourse cannot, by their nature, guarantee that the word or segment of interest will occur in an easily segmentable context. In such circumstances the consistent application of well-specified criteria becomes even more important, to reduce variability in measurement where possible, and ensure that the results obtained are not simply an artefact of the segmentation process.

A single set of criteria was used for locating the word boundaries and segment boundaries. They derive from those which I helped to develop for the ATR/CSTR Speech Database Project (Laver *et al.*, 1988, Laver *et al.*, 1989a and Laver *et al.*, 1989b). Here I summarise the general principles adhered to for each main segment type.

Unless stated otherwise, boundaries were drawn at the sample nearest to the **zero crossing**, and consistently marked at the start of the waveform cycle, i.e. on the steep 'up' stroke (see Figure 5.2).

5.3.1 Stops

Stop consonants are potentially composed of three acoustically distinct events: a period of closure, a release burst, and aspiration (delayed Voice Onset). I shall refer to the burst and aspiration, when taken together, as the **release phase** of the stop. Whether all three acoustic events are realised depends on both the **voicing** of the segment and the **syllabic structure** in which the segment occurs: whilst all three components are likely to be present in syllable-initial stop consonants occurring in stressed syllables, initial stops in weak syllables are

characterised by weak bursts followed by little or no aspiration. (Indeed these bursts may be too weak to be evident in either a spectrographic or time/amplitude waveform display.)

Since closure, burst and aspiration are easily identifiable, each component was segmented when it occurred, on the grounds that the durational values for the stop components could readily be summed to produce a durational score for the whole segment if required, whilst a division in the reverse direction is not possible. Thus maximal information could be preserved and a comparison made of duration reduction in closure versus release phases, should one want to locate more precisely the locus of stop segment compression.

Laver *et al.* (1989a) recommend that the onset of stop consonant **closure** be marked at the point at which energy in the region of F2 and of higher formants ceases to be visible on the spectrographic display. This allows for the presence of voicing through part or all of the closure, while excluding transition information in the offset of vowels which precede closure. In the majority of cases this criterion can readily be adhered to. See, for example, the segmentation of /b/ in Figure 5.4.

The one major problem is with the **frication** of some stop segments, i.e. frication noise generated at the point of constriction as opposed to aspiration noise generated at the glottis. One of the manifestations of weakening in stop articulation is failure to make complete closure; this lack of closure leads to frication rather than silence. A stop which is completely fricated can be segmented at the onset and offset of the frication period, which is clearly visible in spectrographic displays, and labelled as "fricated stop". However, a stop with fricated pre-closure before a small but definite closure period is more problematic. The criterion above suggests that the stop ought to be segmented only at the onset of actual closure, but, certainly in the present materials, the frication period clearly belongs to the stop rather than the preceding segment. In these few cases, the fricated onset was included in the overall stop duration measurements, but labelled independently as a separate phase. Thus the label for the fricated /d/ of *walled city*[3] would consist of three elements as in (5.1):

(5.1) /wɔl [d_{fric}] [$d_{closure}$] [$d_{release}$] sɪtɪ/

This is illustrated in Figure 5.3.

The onset of the stop consonant **release** phase coincides with the burst, where there is one, or, where there is no visible burst, the onset of any aspiration. The

[3]first mention from q1ec3

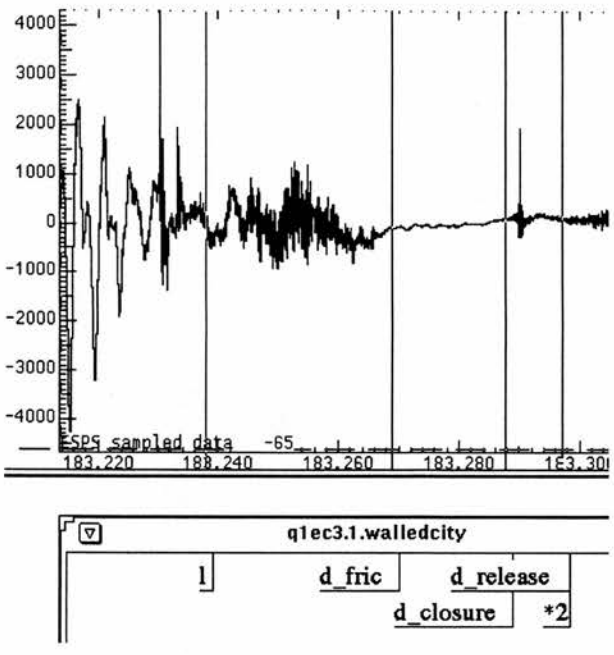


Figure 5.3: Time/amp waveform and label file for first mention of *walled city*; figure illustrates use of labels for different phases of stop articulation: fricated, closure and release. *2 indicates the offset of the word *walled.*

release is taken to include everything up to the onset of the following segment, that is, to the point of the first identifiable glottal pulses for neighbouring voiced segments, and/or the appearance of clear formant structures. Usually this point will coincide with an absence of high frequency 'friction' energy, indicating the end of the burst/aspiration. Where a stop is followed by a strident fricative such as /s/ or /ʃ/, however, it is the relative *increase* in high frequency noise that marks the onset of the continuing context (such as the /s/ in (5.1) above). Since the /d/ preceding /s/ in 5.3 is a voiced stop and not aspirated, there is no confusion over the assignment of the high frequency energy to the fricative. In fact, Laver *et al.* (1989a) observe that stops preceding fricatives and nasals will in general have no identifiable release phase, in which case they suggest that only the closure is marked.

In the case of [stop]+[stop] sequences, as in *old temple*, there is usually no discernible release of the first stop segment. In these situations, Hieronymus *et al.* (1990) recommend dividing the period of silence at the midpoint, with half of the closure period being allocated to each of the two stops. Where a release was evident, as occasionally happened in citation form productions, the stop was segmented according to the standard criteria above. For the purposes of excerpting material for experiments on intelligibility I adhered to Hieronymus *et al.* (1990)'s recommendation. However, for analyses of word duration I used the combined

(closure$_{stop1}$ + closure$_{stop2}$) duration (see Chapter 8, Section 8.3.2).

In utterance-initial or 'post-pausal' location the onset of closure for stop consonants is difficult to differentiate from preceding silence. Because citation form productions are necessarily post-pausal, separated as they are by periods of silence, *all* tokens of name-initial stops, such as the /k/ but not the /b/ of *Crane Bay* and the /d/ of *disused monastery*, were measured from the onset of the **release**. Thus tokens of *Crane* taken from context where there was a clear closure onset were also measured from the release point. Only in this way can the duration of /k/ in *Crane* spoken in context be compared with the duration of /k/ in *Crane* taken from the citation form. Although this results in the loss of information about the duration of the stop closure in first and second mentions of the landmark name in context, it was felt that this was acceptable given that no direct comparison between tokens in context was being made. In all cases word and segment duration was compared against a matched citation form, with the difference from citation form being the dependent variable in subsequent analyses.

Utterance-final stop consonants present a different problem. Laver *et al.* (1989a) recommend that:

> *"the release phase should only include the burst and not the pre-silence weak friction which is sometimes present since the duration of these breathy offsets is highly variable and contains no phonemically relevant information." (Laver* et al., *1989a, page 13)*

On this basis, the offset of utterance-final segments was in general segmented **as early as possible**, consistent with the absence of significant energy in the region of F2 and above, and the end of any periodic patterning in the time/amplitude waveform. The material excluded in this way contains only a noisy release of breath.

5.3.2 Fricatives

The segmentation of most of the fricatives in English is relatively straight forward, since the high frequency noise excitation is quite distinctive. There is usually little difficulty in identifying the four strongest fricatives /s, ʃ, z, ʒ/. Locating the start of landmark names such as *saloon, savannah, shelter* and *shore* was rarely problematic.

Place of articulation can generally be differentiated on the basis of spectral shape, with the major energy distribution for /ʃ/ and /ʒ/ starting at a lower frequency

than for /s/ and /z/, for example. This helps to locate boundaries in sequences such as *collapsed shelter* and *crashed spaceship* when the /t/ has been fully deleted by a process of apocope.

Whilst the voiceless fricatives, /f/ and /θ/, are also relatively easy to locate, the voiced weak fricatives, /v/ and /ð/, may sometimes present a problem, especially in the realisation of function words such as *the* and *of*. The very short duration of such weak fricatives makes them difficult to label unambiguously. However, they usually show an overall drop in energy in the formants and some evidence of frication, if not spectrally, then on the periodicity of the waveform. The same holds true for the weak glottal fricative /h/, which occurs in the landmark names *haystack, hideout, hostel, house* and *hut.*

In cases where there is a short period of apparent 'silence' between a vowel and a following fricative, the silence was assigned to the fricative (Laver *et al.*, 1989a), with the fricative onset being taken as the offset of the preceding vowel, that is, the cessation of visible energy in the region of F2 and above.

5.3.3 Affricates

The segmentation criteria for affricates combines those for fricatives and non phrase-final stops. Thus, closure boundaries were determined following the guidelines for stop consonant closures, while the offset of the frication portion of the affricate was located at the cessation of high frequency energy. Although it would be possible to segment affricates sub-phonemically into separate closure and frication phases, this was not done, and a single duration for the segment as a whole was all that was measured. Landmark names involving affricates at onset included *chapel* and *giraffes* with *village* and *forge* having affricates at offset.

5.3.4 Nasals

In general, determining the appropriate boundaries for nasal consonants is relatively easy. The onset of the oral closure made during the articulation of a nasal stop consonant involves a clear, distinct fall in energy above 500Hz relative to the preceding segment (Laver *et al.*, 1989a). The offset is similarly characterised by an immediate increase in energy above 500Hz, except, of course, when preceding an oral stop consonant, and the nasal is thus adjacent to silence, corresponding to stop closure. When the nasal is intervocalic, there may also be evidence of **discontinuity** in formant frequency, in addition to reduction in formant amplitude

during the nasal (Hieronymus *et al.*, 1990).

In sequences of [nasal] + [nasal] where the two nasal segments are produced at different places of articulation, there is usually a discernible boundary between the two nasals as a result of a visible shift in formant frequency location. For geminates, on the other hand, there is no evidence other than duration to indicate the existence of two rather than one segment.

The nasals studied in Chapter 7 were exclusively word-final lexical alveolars preceding either labial or velar stop consonants. The onset of the nasal always followed a vowel, unless the nasal was realised as a syllabic segment (such as in words like *seven* and *golden*), while the offset preceded the silence of a stop closure. The nasal onset was therefore taken to be the start of the pitch period immediately following the discontinuity that marked the transition from vowel to nasal murmur[4]. Nasal offset was sometimes harder to locate. In some cases, there was clear evidence of a nasal release, represented by a visible burst of energy. This is more common for /n/ and /ŋ/ than for /m/ (Hieronymus *et al.*, 1990). These bursts were segmented separately so that they could be analysed independently if desired. In cases where there was no burst, segmentation boundaries were drawn at the point at which the nasal murmur waveform started to lose its regular pattern. This point usually coincided with a reduction from low to minimal amplitude, which was interpreted as marking the silence of the oral stop closure.

5.3.5 Liquids and glides

Although this thesis work focuses on the durational aspects of stop consonant and nasal articulations, the accurate segmentation of liquids and glides was important for the extraction and duration measurement of words used in the intelligibility experiments, such as *waterfall, warehouse, youth-hostel, yacht-club, ravine, rock, lake, lagoon* and *lighthouse.*

The glides /w, j/ and the liquids /r, l/ are probably the most problematic phonemes to segment. In the context of vowel neighbours, /r, w, j/ are characterised by large formant transition regions, with little or no steady-state that can be assigned clearly to the consonant as opposed to the adjacent vowel. For the sake of consistency, segment boundaries were drawn at the *midpoint* of the

[4]This was usually easy to locate on the time-amplitude waveform, since nasal murmurs tend to be characterised by a rather simple waveform pattern, compared with the complex waveform of a vowel segment.

transition into and out of these problematic consonants (Laver *et al.*, 1989b).

While clear /l/ provides relatively unambiguous cues to boundary location, with a discontinuity in F1 and F2 not unlike that for nasals, locating the boundaries of dark /l/ is more difficult, and, in the context of vowels like /u/, /ʊ/ and /o/ (the Scots equivalent of /əʊ/ in words like *go*), may be close to impossible. This presented problems when measuring stressed vowel duration in words like *old* and *gold* where there is no obvious boundary between the vowel and /l/ segment. For the sake of consistency, I decided to omit the boundary altogether and compare, instead, the durations of the combined /o/+/l/ segments; this seemed acceptable on the grounds that the measurement of interest was the reduction in duration between citation and spontaneous productions of tokens of the same word form, rather than the difference in duration between an /o/ vowel from *gold* and an /o/ vowel from *phone*, for example.

5.3.6 Vowels

As Laver *et al.* state (1989a:17), "segmentation criteria for vowels adjacent to consonants are implicit in what has been said above about consonant boundary placement". In the case of vowel sequences, as in an RP production of the word *iron* (from the landmark name *iron bridge*), pronounced [aɪən], segmentation may depend on formant changes (e.g. movement away from a target maximum) and/or variations in amplitude, effected by, for example, changes in lip-rounding. Only three such cases occurred in the Map Task data analysed in this thesis: *diamond, iron, lion*; of these, *diamond* was rarely pronounced [daɪəmənd], while roughly half the productions of *iron* involved a tapped /r/, as in [aɪrən].

5.3.7 Examples

To illustrate the application of these segmentation criteria to the Map Task data used in the experiments to be described, I present below a few examples. Each example is associated with a short description detailing any problems encountered, and drawing attention to particular aspects of the segmentation that bear on the research. The transcription is a machine readable version based on the CELEX SAM-PA scheme (see Appendix B) but without the length marking for vowels. The onset of the first word is marked ['1]. The symbol [!] denotes stop release, which may include burst and/or aspiration.

Example One: *saloon bar* (q5e.cit) Figure 5.4 illustrates a carefully articulated citation form production of *saloon bar* which presents few segmentation difficulties. The discontinuity of F1 and F2 associated with the clear /l/ is easily identifiable, as is the energy loss associated with the nasal. There is evidence of voicing during the closure period of the /b/ at the onset of *bar*, followed by a short burst. The rest of *bar* was not segmented (i.e. no boundary was drawn between the vowel and /r/ segment) since only the word onset and offset was required for this word.

Example Two: *carved stones* (q3n.cit) Figure 5.5 illustrates the need to segment stop-initial words, such as *carved*, from the onset of the burst, because of the preceding silence in citation form productions. The /d/ of *carved* is clearly not deleted in this production, with evidence of a closure period followed by short burst. There is no evidence of voicing, however[5]. There is evidence to suggest that the /o/ vowel of *stones* is nasalised, but the onset of the nasal itself is still identifiable by the discontinuity in both the time/amp waveform and spectrographic display. The end of *stones* illustrates the problem of pre-pausal breathiness in the waveform. The offset of the word is marked as soon as the spectrographic display shows no significant energy in the region of F2 and above. When the segment between the offset of *stones* and the dotted line is played, no evidence remains of the preceding fricative and the segment simply sounds like rather breathy exhalation.

Example Three: *concealed hideout* (q8e.tok2) The final example, Figure 5.6, presents a rather different picture. While the earlier examples were taken from clearly articulated citation forms, this figure depicts a far 'muckier' token: a repeated mention of *concealed hideout*, where the guidance to segmentation offered by either time/amp waveform or spectrogram is poor. Even the /s/ of *concealed* is problematic: the voicing striations make the boundary between /s/ and /i/ rather difficult to locate. There is no evidence of any /d/ at the end of *concealed.* The section of speech in the waveform and spectrogram windows which is marked by the unlabelled vertical line[6] sounds more /h/-like than /d/-like, and is definitely the start of *hideout* and not the end of *concealed.*

[5]The main difference between the devoiced /d/ of *carved* and the voiceless /t/ of *stones* is in the duration and spectral characteristics of the release.

[6]This line was left over from a play-back routine run from the time/amp window, and is not a segmentation line.

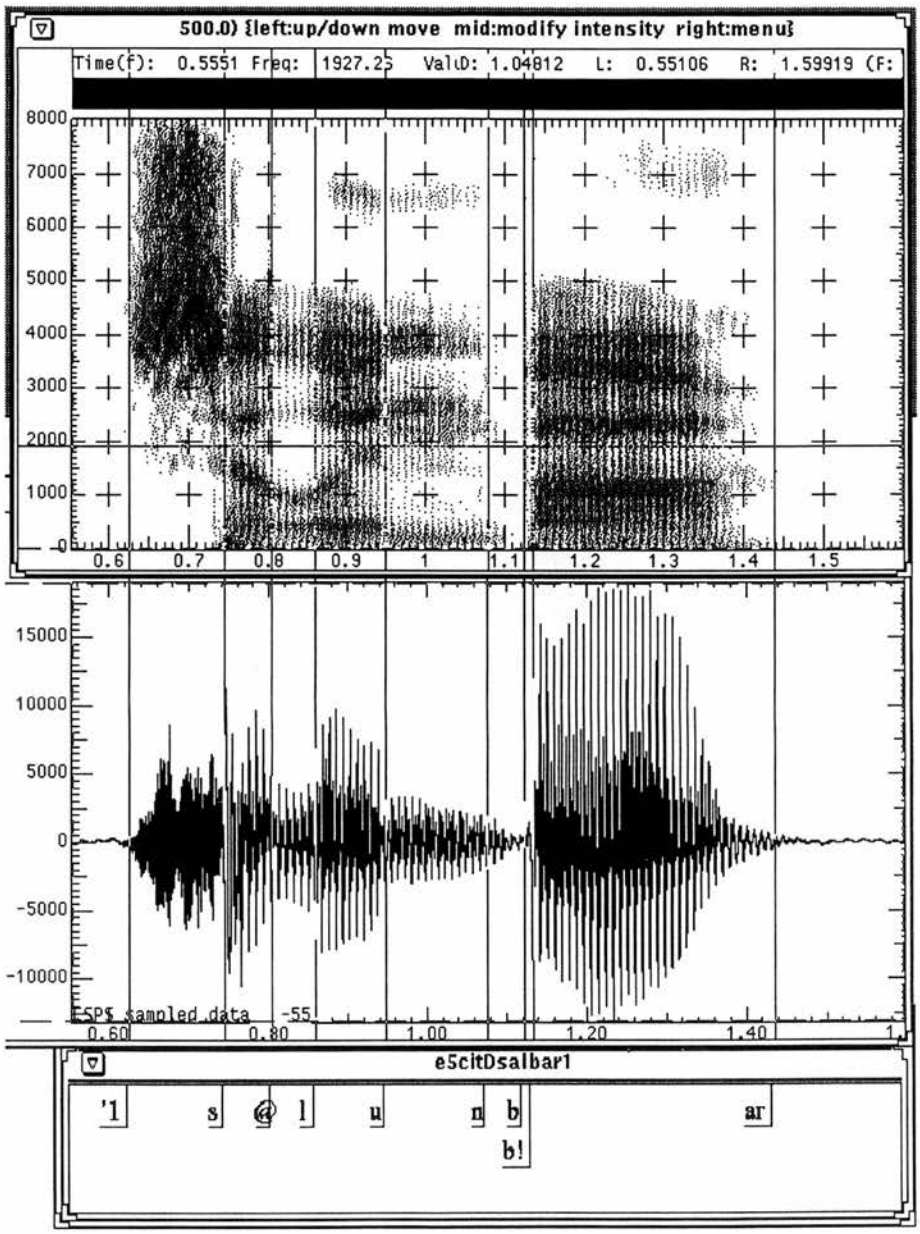


Figure 5.4: Spectrogram and time/amp waveform illustrating segmentation boundaries for a citation form production of *saloon bar*.

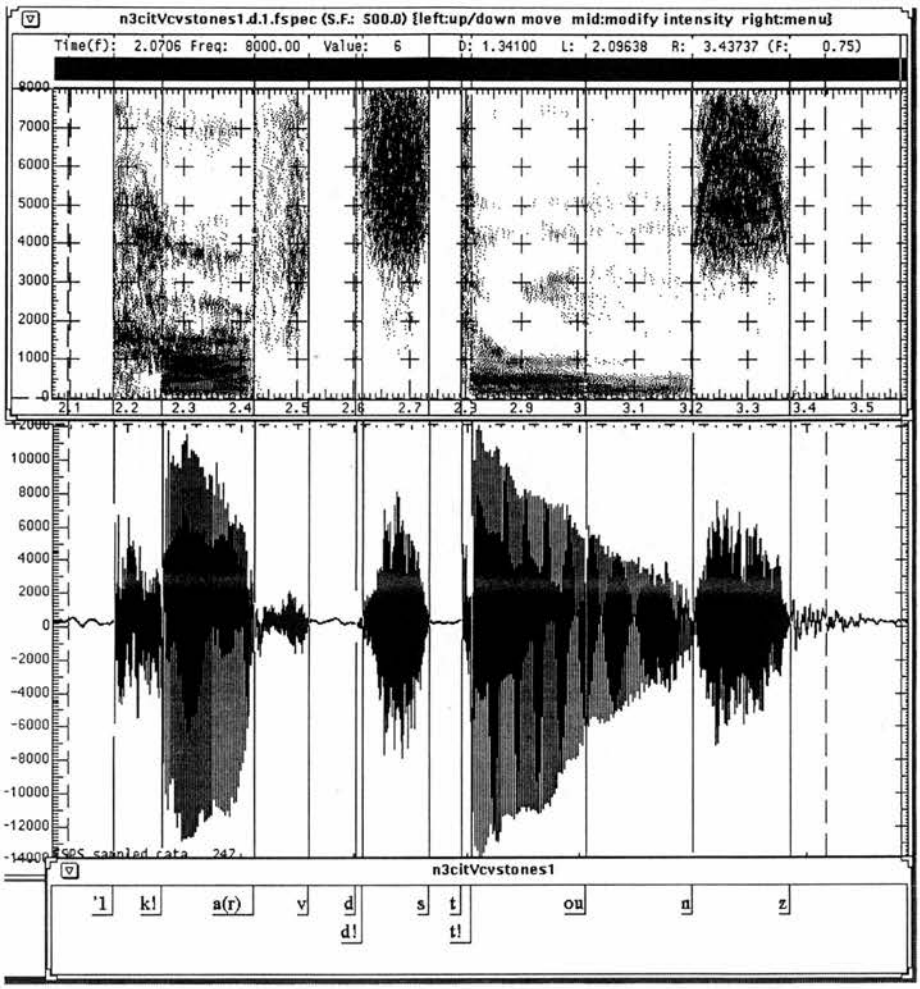


Figure 5.5: Spectrogram and time/amp waveform illustrating segmentation boundaries for a citation form production of *carved stones*.

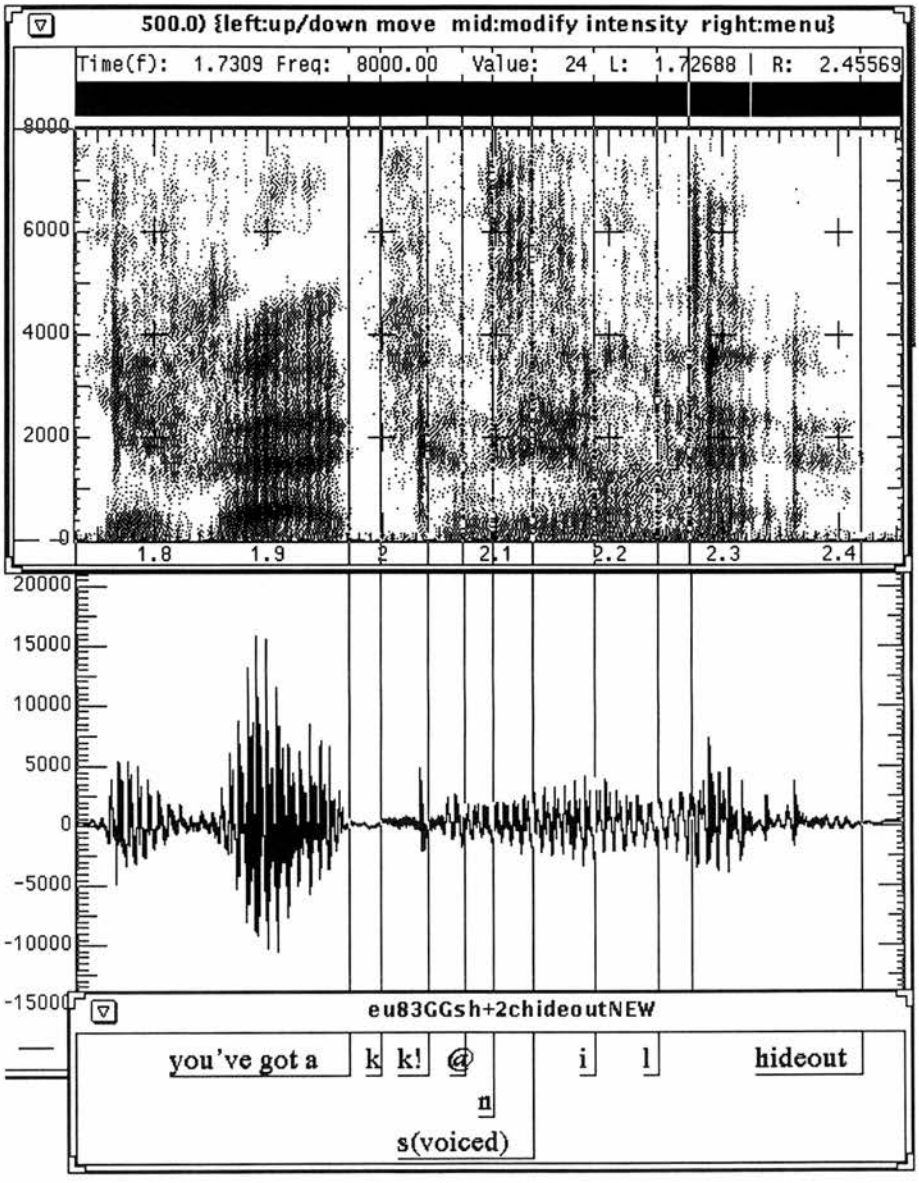


Figure 5.6: Spectrogram and time/amp waveform illustrating segmentation boundaries for a repeated mention of *concealed hideout*.

Chapter 6

Discourse Repetition Effects on Intelligibility

6.1 Introduction

Lindblom's H & H theory claims that the level of hyper- or hypo-articulation in the speech signal depends on the availability of signal-independent information, or contextual support: the greater the degree of contextual support to aid recognition, the more likely speakers are to hypo-articulate. In Chapter 4 I introduced the distinction between 'Given' and 'New' information, and discussed some of the linguistic means by which this difference can be conveyed. I detailed how co-referential repetitions – references to Given information – have been shown to be shorter in duration, and less intelligible than corresponding tokens that introduce the entities into the discourse – references to New information. Further, while these introductory first mentions are almost always accented, repetitions are frequently associated with deaccenting.

In this chapter, I describe a series of intelligibility experiments which were conducted to explore the extent to which the repetition effect discussed in Chapter 4 is **co-operative**. The H & H theory presents speakers as cooperative participants in dialogue, adjusting articulatory effort according to a gross running account of listener needs (Lindblom, 1990a, page 405). However, there is evidence to suggest that degraded tokens sometimes occur on occasions when the listener is unable to recover the conditioning information (see Section 4.9). For example, adults reduce intelligibility when repeating themselves to children who failed to attend to the earlier utterance (Bard and Anderson, 1994); speakers also produce less intelligible tokens in recorded dictations, where the original utterance is erased by the repetition (Bard *et al.*, 1989). Reduced intelligibility under these circumstances

is presumably less than cooperative.

To explore the issue of cooperation in relation to the intelligibility loss associated with coreferential repetition, a series of experiments was devised that explored the effects of repetition on intelligibility when the availability of information about the referent is varied for speaker and/or listener. The experiments were designed to answer the following questions:

1. Whose mention counts as a previous mention? Does each speaker maintain a separate account of entities to which they have referred, or do speakers establish and maintain a common record, where introduction by either dialogue participant is sufficient for any subsequent reference to be reduced?

2. To whom must an object be visible to be treated as Given? Do speakers mitigate intelligibility loss when their listener cannot see what is being referred to? Do repeaters mitigate intelligibility loss when they themselves cannot see what they are referring to?

3. If previous mention makes an entity Given, who must know about that mention? Speaker or listener? Is the speaker's record of textually-evoked Given items annotated with the identity of the listener?

In all cases the critical measure is the relative reduction in intelligibility – or **intelligibility loss** – of the tokens excerpted from unscripted dialogue, compared with a base-form measure for that speaker, the citation form. By comparing the intelligibility of spontaneous tokens with matched citation forms it is possible to control for variation in both speakers and materials. Clearly it would be inappropriate to compare speaker A's introductory mention of *volcano* directly with speaker B's repetition of *volcano*, since it would be impossible to tell whether any resulting intelligibility loss was a consequence of repeated mention or of differences in speech style between A and B: speaker B might always produce tokens that are less intelligible than A's clearer productions. However, if speaker B's token of *volcano* (a second mention) is much less intelligible than his own citation form, while A's token of *volcano* (a first mention) is only marginally less intelligible than her citation form, then we can begin to suspect an effect of repetition. In the same way, it is desirable to differentiate between spontaneous tokens which are hard to identify because they have undergone significant reduction compared with their more carefully articulated citation form, from tokens which are hard to identify because they are simply difficult to recognise under all and any circumstances, i.e. even in clear citation productions.

The results of this work were presented at the XIIIth International Congress of Phonetic Sciences at Stockholm, Sweden, in August 1995 (Bard, Sotillo, Anderson, Doherty-Sneddon, and Newlands, 1995).

6.2 Experiment One: Same vs. Different Speaker Repetition

Is referring to an entity that you yourself have introduced the same as referring to an entity that your partner has already mentioned? The answer to this question lies in the nature of the discourse model built by the speaker: if participants in a dialogue attempt to construct a common model of joint activity, then reference to an entity which her partner has introduced is no different from reference to an entity introduced by the speaker herself. If, however, both dialogue participants are building separate, independent models of the discourse, then reference to an entity not introduced by the current speaker will function not as a coreferential repetition but as a first, or introductory, mention.

It should be possible to test which of these two alternative views holds, by comparing self- with other-repetition. If speakers cooperate in building a shared discourse model then other-repetitions will behave like self-repetitions; if, however, speakers construct separate models, then a decrease in intelligibility will be found only for self-repetitions.

6.2.1 Materials

Data was taken from the HCRC Map Task Corpus (Anderson *et al.*, 1991). Landmark names can be introduced by either the Instruction Giver or the Instruction Follower, while repetitions can be by the same speaker who introduced the item or a different speaker. Thus the corpus contains examples from all four cells in Table 6.1.

All 128 dialogues were examined and the first and second reference to each landmark – or feature – located in the transcription. Every feature in each dialogue was coded for the following information:

Sharedness Features were Shared (Sh) or Unshared; if Unshared, they were coded for whether they were present on the Giver's (UG) or the Follower's (UF) map.

		Introducer	
		Giver	*Follower*
Repeater	*Giver*	GG	FG
	Follower	GF	FF

GG and FF are **self**-repetitions;
GF and FG are **other**-repetitions.

Table 6.1: Options for self- and other-repetition by Giver and Follower in the Map Task Corpus

Referring Expression Second mentions were coded according to whether speakers used:

- the literal referring expression offered by the landmark label (L) (such as *the extinct volcano*);
- a reduced referring expression (R) (such as *the volcano*);
- a pronominal expression (P) (such as *it* or *that*);
- something else, or 'Other' (O) (such as *the hill*).

First mentions were assumed to be full referring expressions; exceptions were noted.

Identity of Speaker for each mention First/Second mention pairs were coded as GG, GF, FG or FF, where G stands for Giver, F for Follower, and the ordering reflects the speaker ID for first and second mention.

Feedback The existence of a feature may be explicitly denied (or acknowledged) between the introductory and repeated mention. This was coded as

- '1' if the feature was denied;
- '+' if the feature was acknowledged;
- '–' in the absence of an informative response.

Additional information This allowed for the opportunity to record any other pertinent information.

The coding for dialogue q4ec1 from which the extract in Section 5.2.6.1 was taken, is presented in Table 6.2 by way of illustration. Codes were scored as 1 or 0, where 1 indicates a positive score for that category.

For the studies reported here, only **literal mentions** of feature names were used. Any first or second mentions which involved reduced referring expressions were

landmark_name (q4ec1)	Sharedness Sh/UG/UF	RefExpr L/R/P/O	SpeakerID GG/GF/FF/FG	Deny	
extinct_volcano	1 0 0	1 0 0 0	1 0 0 0	-	
tribal_settlement	0 1 0	1 0 0 0	0 1 0 0	-	
rope_bridge	1 0 0	1 0 0 0	0 0 0 1	-	
machete	0 0 1	1 0 0 0	0 0 0 1	1	
collapsed_shelter	1 0 0	0 0 1 0	1 0 0 0	-	pro (same_turn)
crevasse	0 0 1	0 0 1 0	0 0 0 1	1	dctc
saxon_barn	1 0 0	1 0 0 0	1 0 0 0	+	
pelicans	0 1 0	(one mention only - G)		1	
white_mountain	1 0 0	1 0 0 0	0 1 0 0	-	
golden_beach_2:1_r	0 1 0	0 0 0 1	1 0 0 0	-	
golden_beach_2:1_i	1 0 0	0 0 0 1	0 0 0 1	-	
IG: to the golden beach at the top left-hand corner					
IF: golden beach is in the top right-hand corner					
IG: left-hand corner, sorry. Oh, there's two golden beaches					
slate_mountain	1 0 0	1 0 0 0	0 0 0 1	-	
submerged_rocks	0 0 1	(not mentioned)			
secret_valley	1 0 0	1 0 0 0	1 0 0 0	-	

Table 6.2: Coding of first and second mentions of Map Task features according to sharedness, literal use of landmark label, identity of speakers who uttered first and second mention, and feedback about feature availability: data from dialogue q4ec1

excluded in order to control as best we could for both the neighbouring phonetic context and the metrical structure of the stimulus word.

This restriction might have posed difficulties. Hawkins and Warren (1994) observe that repetition in natural conversation results in frequent pronoun use, and claim that, with the exception of proper nouns, there is little opportunity to repeat words simply as Old information. They argue that when a word *is* repeated in conversation, it is usually for reasons that do not favour deaccentuation, such as the need for contrastive stress, and that these tokens are therefore atypical. It is clear from the distribution of referring expressions in the Map Task data used here, however, that pronominalisation was relatively infrequent, with literal second mentions being far from abnormal. As Table 6.3[1] shows, although the incidence of pronominalisation increases with repetition, over two-thirds of second mentions are either definite or indefinite literal mentions. We suggest that the relatively

Referring	**Mention**	
Expression	First	Second
indefinite	48.7%	19.3%
definite	50.9%	49.9%
deictic	00.0%	14.3%
pronoun	00.4%	16.5%

Table 6.3: Distribution of types of referring expression for first (N=631) and second (N=607) mentions of landmark names in the Map Task

high incidence of full referring expressions found in our data was a consequence of the difficulty in resolving pronominal and deictic anaphora in contexts where the expression '*it*' could refer to any one of a number of landmark names. Further, the incidence of contrastive stress in the data appears to be relatively low, despite the deliberate incorporation of contrasting features (*gold mine/diamond mine*) into the map design. For these reasons we had no reservations about using literal repetitions.

All repetitions had to be both **sequential** – i.e. a genuine second mention, not a third or fourth mention – and **coreferential**: use of 2:1 features (such as *golden beach*, which appears on the map pairs in Figure 5.1 above) was restricted to cases where both first and second mention referred to the same landmark in the same location, with no intervening references to the landmark of the same name located elsewhere. This condition excluded the use of *golden beach* from dialogue

[1] data supplied by M. Aylett

q4ec1, since the second mention of *golden beach* is non-coreferential with the first (see Table 6.2).

The excerpt in (6.1) (from q4ec1) illustrates coreferential self-repetition by the Instruction Giver (GG). The relevant referring expressions are in italics with the order of mention in brackets:

(6.1)
IG: Start at *the extinct volcano* (1), and go down round the tribal settlement. And then
IF: Whereabouts is the tribal settlement?
IG: It's at the bottom. It's to the left of *the extinct volcano* (2).
IF: Right. How far?

The same extract also contains a coreferential other-repetition where the feature is introduced by the Instruction Giver and then referred to by the Instruction Follower (GF). This is illustrated in (6.2).

(6.2)
IG: Start at the extinct volcano, and go down round *the tribal settlement* (1). And then
IF: Whereabouts is *the tribal settlement* (2)?
IG: *It* (3)'s at the bottom. *It* (4)'s to the left of the extinct volcano.
IF: Right. How far?

For the first experiment material was selected from one half of the corpus: Quads 3, 4, 7 and 8, both with and without eye-contact. These Quads were the first to undergo several levels of linguistic coding and therefore there was more information available with respect to speaker and listener behaviour to support the interpretation of our results.

First and second mentions were taken from first givings only, i.e. dialogues 1 to 4, except when they were same-speaker Follower repetitions (FF) in which case tokens from dialogues 5 to 8 were also included. This ensured that all first mentions were genuine first encounters with the landmark in question. The Given/New status of first mentions by Instruction Givers on the second occasion they see the map was *a priori* unknown: we addressed this empirical question directly in Experiment Four below.

First and second mentions were also restricted to **shared** features, since the self/other repetition variable is confounded in unshared features with whose unshared feature is being discussed. Because speakers only introduce a feature which occurs on their own map, i.e. *their* unshared feature[2], self-repetitions of unshared features are always self-repetitions of one's own feature while other-repetitions of unshared features are invariably other-repetitions of the partner's feature.

In all cases subjects were presented with single word tokens. These were either single word landmark names, such as *pelicans* and *crevasse*, or words taken from longer referring expressions, such as *Saxon*, taken from *Saxon barn*. In the latter case, the use of initial and final components of landmark names was balanced across conditions to avoid any phrase-final duration bias. For the same reason, the distribution of single word feature names and words from longer names was balanced across conditions.

6.2.2 Design

A total of 96 suitable word types were found: 48 self-repetitions (GG = 28, FF = 20) and 48 other-repetitions (GF = 19, FG = 29). Given that it was not desirable to make any subject try to recognise two tokens of the same word type, it was only feasible to balance word types across repeater conditions by splitting the material and presenting it to twice as many subjects. Thus each subject heard 48 items: 24 self-repetitions and 24 other-repetitions, where each was a different word type. It was not possible to achieve a perfect word match across repeater condition: while 18 words appeared as both self- and other-repetitions, the remaining 30 word types were matched as well as possible on syllable length and word frequency. Where feasible, speakers were balanced across the two conditions so that speakers who produced both self- and other-repeater tokens were selected in preference to speakers who produced only self- or other-repetitions, in order to reduce possible effects of individual speaker variability.

For each word type there were four separate tokens:

- Token1: introductory mention from the unscripted dialogue
- Citation1: citation form by speaker of Token1

[2]This is true of first givings, i.e. dialogues 1 to 4, although it does not necessarily hold for all second givings: Givers sometimes introduce a landmark which their previous Follower had mentioned in the earlier dialogue. As data was selected from first givings only, the exceptions in dialogues 5 to 8 are irrelevant.

- Token2: second mention from the unscripted dialogue
- Citation2: citation form by speaker of Token2

In the case of self-repetition, Citation1 and Citation2 are the same word token. Thus there were 48 items, by 2 repeater conditions (Self/Other), by 2 speech form conditions (Token/Citation) by 2 mention conditions (First/Second), resulting in a total of 384 tokens.

6.2.3 Procedure

The relevant utterances were located on the original Digital Audio Tape (DAT) recordings, and digitised at a sampling rate of 16kHz using an Ariel s32c DSP card via a Proport Audio/Digital box. The resulting speech files were inspected using the Entropic Signal Processing System (ESPS) WAVES software package on a Sun Sparc station. Word boundaries were determined using a combination of time/amplitude waveforms, spectrograms and auditory playback. Boundaries were placed at zero crossings, and the isolated word played to check for any problems introduced by the segmentation procedure. The criteria adopted for locating segment boundaries are described in Chapter 5.

Many of the words were highly frequent and/or polysyllabic, both factors which correlate with intelligibility, so the excerpted tokens were masked by noise in order to avoid ceiling effects. This was achieved by multiplying the original speech data file by a 16KHz 5-second file of random noise (where all sample values were in the range 0.5 to 1.5). The amplitude of the resulting stimulus was related to that of the original speech data file, and the data points retained the same sign as the sampled data values they replaced, but were scaled differently. The distortion varied from reducing sample points by one half, to increasing sample points by 1.5 times their original value[3].

The noise overlayed speech files were recorded on to Digital Audio Tape with an Inter Stimulus Interval (ISI) of 8 seconds. Word tokens were allocated to 8 different presentation tapes according to a Latin square design. The order of presentation of items was randomised, with the same random order of words adopted for each presentation tape, to help avoid errors in collating the results.

[3]The concept of a signal-to-noise ratio is inappropriate in this context since it applies only to signals where the original data and noise have been *added* together. The ratio is then a reflection of the relative sizes of the two. Since the signal and noise were multiplied on this occasion the situation is quite different, and a signal-to-noise ratio here would be meaningless.

There were four examples, after which subjects were given an opportunity to ask questions. The 48 test stimuli were then presented without a further break.

6.2.4 Subjects

Subjects were native speakers of English with no known hearing impediment. They were all undergraduate students at the University of Glasgow. Many were themselves speakers of West Coast Scots; the remainder had at least encountered the regional variation in accent represented in the speech stimuli as a result of living and studying in Glasgow. Each tape was played to ten subjects who listened to the stimuli over headphones in individual sound-proofed booths. Responses were typed directly into the computer using the Word text editor on an Apple Macintosh.

6.2.5 Results

Recognition responses were scored as 'correct' only if they were letter perfect. In a subsequent analysis the criterion for correct recognition was loosened to include obvious typographical errors, and since the two analyses conform on the critical results the more recent (loose) intelligibility values will be reported here.

Intelligibility is defined as **the proportion of subjects who correctly identified the excerpted stimulus**, with values ranging from 0 (failure by all subjects to recognise) to 1 (100% successful recognition). Scores were submitted to ANOVAs both by subjects and by materials, for both **raw intelligibility scores** – the proportion of correct responses – and measures of **intelligibility loss** – the difference in correct identifications between careful citation tokens and spontaneous mentions.

The raw intelligibility scores, found in Table 6.4, were submitted to an Analysis of Variance with intelligibility loss from citation as the dependent variable, repeater (Self/Other) and mention (First/Second) as repeated measures, and word list (One/Two) as a between items grouping factor. Although the effect of mention was significant by subjects and approached significance by items ($F_1(1,72) = 5.90, p < .02$; $F_2(1,80) = 3.26, p = .075$) the intelligibility loss associated with repeated mention was insensitive to the identity of the speaker who introduced the entity (Repeater x Mention: $F_1 < 1; F_2 < 1$). First mentions were less intelligible than their respective citation forms by .150 in self-repetitions and by .146 in other-repetitions, while second mentions were reduced by .231 and .227

for self- and other-repetitions respectively.

		Speech form		
Repeater	**Mention**	Citation	Token	***loss***
Self	First	.650	.500	*.150*
	Second	.610	.379	*.231*
Other	First	.706	.560	*.146*
	Second	.598	.371	*.227*

Table 6.4: Mean intelligibility (0-1) for first and second mentions of both spontaneous tokens and citation forms, where repetition is either by the same speaker who introduced the feature, or by their partner

Self- and other-repetition, then, appear to give equal amounts of intelligibility loss. Either speaker's mention makes the word sufficiently Given for the repeated mention to be less intelligible; speakers do not have to wait until they themselves have uttered the word before they can treat the word as Given. This suggests that rather than maintain an individual record of entities which she has introduced to the dialogue, a speaker creates a common account which includes entities introduced by both herself and her partner.

6.3 Experiment Two: Effects of Landmark Visibility on Repetition I – Can the *Listener* see the Referent?

The next two experiments explored the possible effects of referent availability. Prince (1981) uses the term 'situationally evoked' to refer to entities that are Given by virtue of being visibly present at the time of speaking. Since some of the landmarks in the Map Task Corpus are present only on one map – that is, situationally Given to just one of the dialogue participants – knowledge about the existence and location of features is not always initially shared. It is possible, therefore, to explore the effects of this knowledge imbalance on intelligibility. Specifically, we can ask how important it is that either the listener and/or the speaker can see the entity being referred to.

Clearly, the speaker cannot know, at the time of introducing an entity, whether the entity is shared by her listener, or available only on her own map. Thus we would not expect to find effects of sharedness on first mentions. Once the entity is

introduced into the discourse, however, the listener has an opportunity to provide feedback to the original speaker about whether or not he, too, can see the feature that has just been mentioned.

The question we address first, then, is whether intelligibility is affected by knowledge that a referent is or is not situationally evoked for the listener: does a speaker mitigate intelligibility loss when her listener cannot see what she is referring to?

6.3.1 Materials

To contrast between entities the speaker thinks her listener can and cannot see we need to use **unshared** features occurring only on the speaker's map. Unshared features ought to elicit negative feedback from the listener in the form of a denial of the feature's existence on their map. But the appropriate feedback is not always offered. Although failing to deny the feature is an error on the listener's part, the speaker has no way of knowing this: without feedback, the speaker may assume, mistakenly, that the landmark is in fact common to both maps.

Literal introductions and repetitions by the same speaker of unshared features were classified according to the feedback received between first and second mention. In the **Deny** condition, listeners provided feedback that explicitly denied (correctly) the existence of the feature on their map; in the **No Deny** condition, there was no such feedback. In the latter context, speakers might have assumed that the feature was *apparently shared*, since no evidence to the contrary was provided. If speakers are being helpful and cooperative, they ought to respond to the listener's denial by degrading the repeated token less.

Example (6.3) (from q4ec7) illustrates a self-repetition of *pelicans* which follows a denial from the listener (speaker B). An example of a self-repetition with no denial is presented in (6.4) (taken from q7nc4).

(6.3)
A: Stop, ehm, beside the 's' of 'Saxon'.
B: Okay.
A: And, have you got *pelicans* (1)?
B: No.
A: No. Ehm, go down about three to four centimetres from the Saxon barn vertically downwards.
B: So I'm above the rope bridge?
A: Just a bit above the rope bridge, yeah.
B: Okay.

A: And then go underneath where I've got *pelicans*(2) towards the left-hand side.

In the transcription below, the symbol '/' indicates that the speech was interrupted at this point by the other dialogue participant, but that the speaker who was interrupted continued talking. In other words, the speech was continuous across the break in the orthographic transcription.

(6.4)
B: Right. Okay.
A: And then it ju ...
B: What about *the banana tree*(1)? Does it /
A: Oh.
B: come before *the banana tree*(2)?
A: I don't have *a banana* /
B: [laugh] Right. Okay.
A: *tree*(3). Where's *the banana tree*(4)? Roughly.

6.3.2 Design

There were 60 items by 2 denial conditions (Deny/No-deny) by 3 token conditions (Citation/Token1/Token2), resulting in a total of 360 separate word tokens to be identified. Tokens came from literal mentions of unshared landmarks where the introduction and repetition were by the same speaker. Half of the word types were identical across denial condition; the remaining 30 were matched for word length and frequency where possible.

6.3.3 Subjects and Procedure

The same procedure was followed as that for Experiment One. There were 54 subjects, 9 subjects hearing one of 6 presentation tapes. The subjects were drawn from the same speech community as for Experiment One, but had not participated in the previous experiment.

6.3.4 Results

Table 6.5 contains the raw intelligibility scores. An Analysis of Variance was performed on intelligibility loss (intelligibility of citation form – intelligibility of running speech token) with mention (First/Second) and feedback (Deny/No-deny) as repeated measures, and word list (One/Two) as a grouping factor between items.

		Speech form		
Feedback	**Mention**	Citation	Token	***loss***
Deny	First	.759	.631	*.128*
	Second	.759	.554	*.205*
No-Deny	First	.731	.491	*.240*
	Second	.731	.572	*.159*

Table 6.5: Mean intelligibility (0-1) for same-speaker repetition of unshared features which were or were not denied prior to second mention

There was no overall increase in intelligibility loss from first to second mention (Mention: $F_1 < 1; F_2 < 1$) nor was there a main effect of listener feedback (Deny: $F_1(1, 52) = 1.31, p = .257$; $F_2 < 1$). Of course since a landmark can only be denied once it has been introduced the critical result required to demonstrate speaker sensitivity to the listener's view of the world is an interaction between token and feedback conditions. Although this interaction is significant (Denial x Mention: $F_1(1, 52) = 21.96, p < .0001$; $F_2(1, 116) = 5.30, p < .05$) second tokens were not found to differ in *post hoc* Scheffé tests. Clarity was lost to the same degree when the listener apparently could see the referent (intelligibility loss = .159) and when he said he could not (loss = .205). When the landmark was appropriately denied second tokens were significantly less intelligible than their introductory mention (Scheffé at $p < .05$ by subjects). Essentially, second tokens were equally reduced, regardless of whether the speaker had received explicit information that the listener had no visual access to the referent, that is, the feature was not on their map and therefore not situationally evoked.

The significant interaction arises from a difference in intelligibility between **introductory** mentions in the Deny and the No-deny conditions, that is, prior to the opportunity to provide feedback. The introductory mentions of features that fail to elicit a correct denial response are significantly less intelligible (loss from citation = .24) than first mentions which succeed in eliciting appropriate feedback (loss from citation = .128) (Scheffé at $p < .01$ by subjects). They are also less

intelligible on introduction than when repeated ((Scheffé at $p < .05$ by subjects). I will return to this finding, and discuss its possible significance, in Section 6.7. For the time being, however, it will be put to one side, with the primary conclusion that speakers do *not* mitigate intelligibility loss when they know that their listener cannot see the entity to which they are referring.

6.4 Experiment Three: Effects of Landmark Visibility on Repetition II – Can the *Speaker* see the Referent?

The results from Experiment Two show that speakers are insensitive to their listener's ability to see the feature being discussed. But perhaps the notion of 'situationally evoked' is only relevant for what the **speaker** can see. To test this hypothesis, a comparison was made between repetitions of landmarks that the repeater could see, i.e. shared features, and those that she could not see: unshared features which were on her partner's map only. If seeing the referent is important for the speaker, then repetition of shared features should be more degraded than repetition of unshared features.

Because speakers do not introduce landmarks they cannot see, we necessarily require other-speaker repetitions. We know from Experiment One that self- and other-repetitions result in the same intelligibility loss for shared landmarks, in other words, that other-speaker repetition is degraded when the speaker **can** see the referent. The question we now ask is whether other-speaker repetition is equally degraded when the referent is **not** visible to the repeater.

6.4.1 Materials

Again, individual words were taken from references to map features in the Map Task Corpus, using the full corpus of 128 dialogues. All repetitions were literal mentions by other-speakers, that is, by the partner of the speaker who introduced the landmark. Features were either shared, as in Example (6.5), or unshared, as in (6.6). The map associated with these excerpts (taken from dialogue q4ec1) can be seen in Figure 5.1.

(6.5) B: Right. How far?
A: Ehm, at the opposite side.

B: To the opposite side. Is it underneath *the rope bridge*(1) or to the lef

A: It's underneath *the rope bridge*(2). And then from the tribal settlement go straight up towards *the rope bridge*(3) and over *the rope bridge*(4). Then down three steps and along to above the volcano.

B: Eh, d ... Is down three steps below or above the machete?

A: Ah. The machete's not on my map.

B: Oh.

A: Down three lines.

(6.6)

A: It's underneath the rope bridge. And then from the tribal settlement go straight up towards the rope bridge and over the rope bridge. Then down three steps and along to above the volcano.

B: Eh, d ... Is down three steps below or above *the machete*(1)?

A: Ah. *The machete*(2)'s not on my map.

B: Oh.

A: Down three lines.

6.4.2 Design

There were 48 items by 2 sharedness conditions (Shared/Unshared) by 2 speech form conditions (Token/Citation) by 2 mention conditions (First/Second), giving a total of 384 tokens. All tokens came from literal mentions where the introduction and repetition were by different speakers. Few landmark names appear both as shared and as unshared features, so it was not possible to balance the distribution of word types across sharedness conditions as effectively as in the previous experiments, where the same word form often appeared in both repeater, or both denial conditions.

6.4.3 Subjects and Procedure

The same procedure was followed as that for Experiment One. There were 36 subjects, 9 subjects hearing one of 4 presentation tapes. Subjects were drawn from the same community as for Experiment One but were required not to have participated in the preceding experiments.

6.4.4 Results

See Table 6.6 for the raw intelligibility results. The data was subjected to an ANOVA with intelligibility loss from citation as the dependent variable, mention (First/Second) treated as a repeated measure, and sharedness (Shared/Unshared) as a grouping variable.

		Speech form		
Sharedness	**Mention**	Citation	Token	***loss***
Shared	First	.782	.632	*.150*
	Second	.757	.456	*.301*
Unshared	First	.810	.581	*.229*
	Second	.831	.421	*.410*

Table 6.6: Mean intelligibility (0-1) for other-speaker repetition of shared features which the speaker can see, and unshared features which the speaker cannot see

A strong main effect of mention was found, with second mentions degrading more from citation than first mentions (Mention: $MinF'(1,116) = 10.53$, $p < .005$). There was also a main effect of landmark visibility or sharedness (significant by subjects and approaching significance by materials), with running speech tokens of unshared landmarks differing more from citation form than running speech tokens of shared landmarks (Sharedness: $F_1(1,35) = 18.50, p < .0001; F_2(1,94) = 3.26, p = .0743$). However, sharedness did not interact with the repetition effect (Sharedness x Mention: $F_1 < 1; F_2 < 1$): intelligibility was reduced from first to second mention whether (difference in intelligibility loss from citation = .151) or not (difference in loss from citation = .181) the speaker had visual access to the referent. Since word forms could not be matched across sharedness conditions it is likely that the main effect of sharedness reflects differences in the set of word forms selected.

The important result for this thesis is that in repeating the other speaker's men-

tion of an entity which both can see, the repeater reduces intelligibility no more than in repeating the name of an entity which she cannot see. In other words, once an entity has been 'textually evoked' by previous mention (Prince, 1981) access to supplementary visual information is irrelevant to production.

6.5 Experiment Four: Same Map to New Follower – Introducing or Repeating?

In the final experiment of the series we examine the effects on intelligibility of giving the same map twice. We have established that speakers reduce intelligibility when referring to entities already Given in the discourse, independent of who introduced the entity, and who can see the referent. But what happens when speakers change dialogue partners? When Instruction Givers are presented with the same map as they gave instructions on previously, and asked to direct a new Follower along what will now be a familiar route, how carefully do they articulate the introductory mention of each landmark name?

Evidently, the information on the map is New to the Follower. But the Giver has already seem the map once before and discussed it with somebody else. Is the critical factor, then, that the listener – in this case the new Follower – has not heard the word before, or that the speaker – the Instruction Giver – has said the word in an earlier dialogue? If the former is critical, then introductory tokens in both first and second givings should be equally intelligible. If the critical factor is that the speaker has uttered the word before, then second givings should yield more reduced introductory tokens than first givings. A cooperative Instruction Giver who adheres to the principles of Lindblom's H & H theory of articulatory effort (Lindblom, 1983b, 1990a) ought to introduce each landmark to the second Follower as clearly as she did to her first, since the Giver's strategy ought to be keyed to how much her *Follower* knows. An egocentric Giver, on the other hand, will reduce introductory mentions second time around because they are already Given for her, albeit they are New to her Follower.

6.5.1 Materials and Design

Full literal mentions of landmark names that were introduced by the Instruction Giver in both the first and the second giving of the map were selected. There were 48 items in each of three token conditions (Citation/Token1-1/Token1-2)

where Token1-1 is the first mention in the first map giving, and Token1-2 is the first mention in the second map giving.

6.5.2 Subjects and Procedure

The procedure and subject community were as before, with no subject having been a subject in any of the previous experiments. There were 54 subjects, 9 subjects hearing one of 6 presentation tapes.

6.5.3 Results

The data was subjected to an Analysis of Variance with intelligibility loss as the dependent variable, and both giving (First/Second) and eye-contact[4] (With eye-contact/No eye-contact) as repeated measures.

There was no simple effect of giving the map twice (Giving: $F_1 < 1; F_2 < 1$). Rather, the effect on intelligibility of giving instructions on the same map to a new listener was shown to interact with the availability of eye-contact (Giving x Eye-contact: $MinF'(1,116) = 7.13$, $p < .01$). A *post-hoc* Scheffé test revealed that when eye-contact was restricted (the No eye-contact condition) introductory mentions were significantly less intelligible on second giving (Scheffé at $p < .01$). The figures in Table 6.7 show that when faces are screened (No eye-contact) there is a greater intelligibility loss from citation form for introductory mentions of landmarks on the second occasion the Instruction Giver presents the map (loss = .182), compared with the first encounter with the map (loss = .072).

In dialogues where speakers had access to the visual communication channel the opposite effect is found: intelligibility loss is significantly greater for first givings (loss = .230) than when the Giver presents the map for a second time (loss = .115). Indeed, in line with the findings of Anderson *et al.* (1997), these first introductory mentions were significantly more degraded than initial introductions of the same landmarks in the no eye-contact dialogues (pair-wise comparisons at $p < .05$ or better in Newman-Keuls tests). This lowering of intelligibility for first mentions during first givings with eye-contact is most probably responsible for the lack of an overall giving effect.

The effects of eye-contact on intelligibility are, however, peripheral to the core concern of this thesis. Rather, the result of interest is that when communication

[4]The eye-contact variable was included for the purposes of a separate investigation described in Anderson *et al.* (1997).

Eye-contact	Mention	Speech form Citation	Token	*loss*
No eye-contact	First	.818	.746	*.072*
	Second	.818	.636	*.182*
With eye-contact	First	.808	.578	*.230*
	Second	.808	.693	*.115*

Table 6.7: Mean intelligibility (0-1) for first and second givings of both spontaneous tokens and citation forms, introduced by the Instruction Giver to two different Followers when both Giver and Follower could make eye-contact and when neither could see the other's face

is restricted to the auditory channel second-pass introductions – where the entity is textually evoked for the Giver but not for her new listener – are less intelligible than first-pass introductions where the first mention is New for both speaker and listener. In other words, speakers appear to be insensitive to the change of listener identity: only the speaker herself needs to have witnessed the previous mention for subsequent reference to be reduced.

6.6 Repetition Effects and Accentedness

In Section 4.8.1 I discussed the suggestion (Hawkins and Warren, 1994) that the effect of repetition on intelligibility arises from a difference in accent status, rather than a distinction between Given/New information directly. In this section, therefore, I describe an examination of a subset of the data used for the intelligibility studies above, which tests the claim that there is no difference in the intelligibility of first and second mentions that cannot be accounted for independently by the presence versus absence of sentence accent.

That there is a relation between accent and discourse status is not disputed; there is now a broad consensus that accent, focus and New information are intimately interconnected (Terken, 1985; Nooteboom and Kruyt, 1987; Terken and Nooteboom, 1987; Eefting, 1991; Vallduví and Zacharski, 1994). Hawkins and Warren suggest that "it is this connection of New information with accent that is important, with accent being the crucial influence on word intelligibility" (1994:p494).

The production experiments of Hawkins and Warren, which investigated the intelligibility of words from 'normal conversational speech', showed no difference

in intelligibility between first and second tokens. Words carrying nuclear stress, on the other hand, were significantly more intelligible than words which were not accented. An analysis of the distribution of nuclear accents revealed that almost all of the New items had nuclear stress (93%), as did a large minority of the Given items (46%). No effect of information value (Given/New) was found for the subset of words carrying nuclear stress. Hawkins and Warren argue from their results that sentence stress affects intelligibility more than the simple Given-New distinction. The question to be addressed, then, is whether the same holds true for the data in the intelligibility experiments described above.

6.6.1 Method

Half of the material used in Experiment Two was selected for presentation to four intonationally trained phoneticians to transcribe.

The selected word tokens consisted of 60 different word types in three conditions: citation form, first mention and second mention. Only self-repetitions of unshared features involving full referring expressions were used. In half of the cases the listener had denied the existence of the feature on their map before the speaker mentioned the entity for the second time, while for the other half of the set there had been no such denial.

The four experts were presented with a tape and a paper transcript. Each tape contained all three tokens of every word type presented in a different randomised order. Tokens were presented in context, with experts both hearing and reading the whole utterance. The transcription offered no indication as to which was the primary word of interest. Each utterance was presented three times, with an ISI of 2 seconds between repetitions, and a longer pause (4 seconds) between different utterances. After listening to the utterance experts were requested to mark the transcription according to the following conventions:

- **Nuclear accents** to be marked with a double underline under the accented syllable;
- **Non-nuclear accents** (i.e. pre-nuclear) to be marked with a single underline under the accented syllable;
- **Unaccented** syllables to be left unmarked.

Nuclear accents were defined, in line with Hawkins and Warren (1994), as the **major pitch movement** within an intonational phrase. This will usually be the

right-most accent, that is, the last accent within the tone group. The nuclear accent is frequently associated with maximal pitch height, maximal intensity and/or maximal duration, though not always. Judgements were based solely on the basis of auditory perception; no visual information in the form of pitch traces or waveforms was provided.

6.6.2 Results

Analyses of expert variability showed that there was no significant difference between experts in the distribution of accent judgements ($\chi^2 = 9.56, df = 6, n.s.$). The categories of 'Accented', 'Non-nuclear accented', and 'Unaccented' were used with equal frequency by all four experts.

The only area of significant disagreement between experts was in the accent judgements of citation forms. Given that nuclear accents are characterised as having maximal pitch movement within an intonational phrase, citation forms are necessarily problematic since they are uttered in isolation; judgements of relative pitch movement/height *etc.* will be at best difficult, and, in the case of monosyllables, for example, logically impossible. Since our primary concern is the relative accent status of first and repeated mentions by the same speaker, the problem of assigning accent status to citation forms is not disabling.

Table 6.8 shows the modal judgement of accent (that is, the accent that most experts assigned to the word) for the three different token conditions: first mention (Token1), second mention (Token2) and citation form. The 3-way accent judgements are presented in two forms: '+/−Accent' groups judgements of 'Nuclear' and 'Pre-nuclear' accent together and contrasts them with 'Unaccented'; '+/−Nuclear' groups 'Pre-nuclear' and 'Unaccented' together and contrasts them with 'Nuclear'.

As one might expect, most first mentions are judged accented (42/60 or 70%) and the majority contain nuclear accents (32/60 or 53.3%). Slightly more second than first mentions are judged unaccented (Token2 = 28, Token1 = 18), though over a third of second mentions still have nuclear accent (25/60 = 41.7%). Most citation forms (49/60 or 81.7%) are heard as accented though the degree of accent (nuclear vs. pre-nuclear) varies.

In comparison to the findings of Hawkins and Warren (1994), fewer of the Map Task first mentions have nuclear stress (93% for Hawkins and Warren's data versus 53% for the data here), while the proportion of second mentions with

	Speech form			
Judgement	Citation	Token1	Token2	***totals***
+Accent	49	42	32	*123*
−Accent	11	18	28	*57*
+Nuclear	26	32	25	*83*
−Nuclear	34	28	35	*97*

Table 6.8: Number of word tokens judged as +/−Accented and +/−Nuclear-accented, based on modal judgement of four intonation experts on 180 stimulus words

nuclear stress is much the same for the two datasets (45.6% for Hawkins and Warren and 42% for the Map Task subset).

The larger proportion of unaccented first mentions (18/60 = 30% vs. 4/57 = 7%) in the Map Task data is largely a consequence of the structure of the landmark names: given a map feature like *tribal settlement* the words *tribal* and *settlement* were analysed separately since they were presented individually in the intelligibility studies; in such cases only one word would be expected to have a nuclear accent and so the other will necessarily be at least non-nuclear and possibly unaccented.

Recall how in Section 4.8.1 it was shown that although Eefting (1991) finds no effect of Givenness[5] on the duration of unaccented tokens, she was unable to report about the duration of Given words which were accented, because of a gap in her experimental design. Although Eefting argued that accented Given words are less acceptable (Nooteboom and Kruyt, 1987) and slower to verify (Terken, 1985) than unaccented, she could not claim that accented Given words do not occur. The results reported here suggest that accented Given words may not be as rare or atypical as Eefting maintains.

Having looked at the distribution of accented and unaccented tokens for introductory and repeated mentions, we turn, now, to the relation between accent and intelligibility. Hawkins and Warren found no difference in intelligibility between first and second tokens but did find a significant effect of accent: words carrying nuclear stress were more intelligible than words that were unaccented. If Hawkins and Warren are correct, then we should expect to find a difference in intelligibility only when there is a difference in accent: while deaccenting of second mentions should result in intelligibility loss, there should be no effect of repetition on in-

[5] where Givenness = same token repeated in following phrase

telligibility for tokens which share the same accent category from first to second mention.

Because of the problem introduced by the poorly articulated first mentions that failed to elicit appropriate denials (see Sections 6.3.4 and 6.7), it was deemed inadvisable to test for a relation between intelligibility and accent for these particular materials; the atypically poor first mentions in the No-deny condition mean that the repetition effect on intelligibility holds only for the data that elicited a denial (N=30). Once the data is further divided between tokens that did and did not share the same accent status across mention, the cell sizes become too small for a statistical test to be meaningful. Instead, analyses were run on the data from Experiment One above, for which first and second mentions had, for independent reasons, been prosodically coded using a version of ToBI adapted for Glasgow intonation (Mayo *et al.*, 1997). An Analysis of Variance run on intelligibility loss from citation, with mention (First/Second) as a repeated measure, and change in prosody[6] as a between items grouping factor (N=71 [no change], N=24 [change]), revealed that prosodic change was unable to predict a significant amount of the variance ($F_2(1, 91) < 1, n.s.$).[7] Second mentions underwent significantly greater intelligibility loss from citation form independent of whether they were deaccented (= change) or not (= no change).

6.6.3 Conclusion

There is no substantive evidence to suggest that the repetition effects reported in this chapter arise simply from a difference in accent status between first and second mentions. The large number of accented second tokens argue against such a conclusion: for the Map Task data at least, the incidence of deaccenting from first to second mention is relatively small. Couple with this the number of first mentions that do not carry nuclear accent as a consequence of occurring in long Noun Phrases and it is unlikely that accent alone will be able to explain the intelligibility loss associated with repeated mention. Further work on the distribution of accents amongst first and second mentions (Aylett, personal communication) reinforces this view, with over 75% of the second mentions used in Experiment One being accented. Further, an ANOVA on intelligibility loss reveals no main effect of prosodic change, nor any interaction between prosody and other variables. The repetition effect, however, remains intact. There is clearly more to

[6] where change was in pitch accent and/or boundary type

[7] The assistance of M. Aylett in running this ANOVA is gratefully acknowledged

intelligibility loss than a simple process of deaccenting.

6.7 Discussion

The intelligibility of word tokens produced in spontaneous dialogue appears to be strongly affected by textual evocation (Prince, 1981): entities that are present in a speaker's record of material that has been mentioned in the discourse are referred to with less intelligible tokens than entities that are mentioned for the first time, and are therefore new to the discourse. Once represented in the discourse model, an entity is named by more degraded word tokens *regardless* of any other speaker or listener knowledge about the earlier mention or the entity. Speakers do not have to wait until they've uttered the word themselves before they can treat the word as Given (Experiment One), nor do they need to be able to see the referent (Experiment Three). Similarly, knowledge that the listener may not have visual access to the entity named does not affect the intelligibility of a word's second mention (Experiment Two).

The loss of intelligibility associated with repeated mention of textually evoked – or Given – entities is, of course, predicted by Lindblom's H & H theory, which claims that speakers reduce articulatory effort when they believe their listener has appropriate contextual information to aid recognition:

> *"Talkers realise phonetic segments in production and phonetic structure is specified in the acoustic signal, only in so far as explicit signal information is needed to supplement implicit contextual listener 'knowledge' " (Lindblom and MacNeilage, 1986, page 130).*

Clearly the representation of an entity in a model of material previously mentioned in the discourse contributes to the signal-complementary sources of information available to the listener to help decode the acoustic input. Recall that there is a strong relation between the cognitive status of a referent and the linguistic choices made by a speaker in referring to it: the more 'activated' or 'accessible' the referent, in terms of the cognitive representation of the discourse, the less informative, more ambiguous, and more reduced the linguistic form used by the speaker to refer to it (Prince, 1981; Ariel, 1991; Gundel *et al.*, 1993). Referents that are 'in focus' (Gundel *et al.*, 1993) can be referred to with maximally reduced expressions. Indeed, it is this reduction which itself conveys to the listener the informational redundancy: the degree of attenuation functions as a marker of accessibility. In other words, a reduced token signals to the listener that the

referent ought to be amongst the most salient in his discourse representation.

Thus far, the empirical results appear to support a Lindblom-like approach to articulatory effort: speakers can afford to reduce articulatory effort in subsequent reference to entities previously mentioned, since listeners can access their model of the discourse to retrieve the appropriate referent, thereby constraining the potential search path.

However, the results above also demonstrate that speakers generate a rather 'thin', skeletal representation of discourse: there appears to be no annotation within the representation of *who* said what, nor, indeed, of who can *see* what. There is no effect of speaker identity: self- and other-repetition results in the same degree of intelligibility loss; and there is no effect of information feedback with respect to visual accessibility: repeated reference to a denied entity is as reduced as reference to an entity that has not been denied. Clearly it is sufficient simply for an entity to have been mentioned before, for the speaker of a subsequent token to reduce articulatory effort.

This finding suggests there are certain limits on the information that the speaker takes account of: while speakers maintain a basic model of what has been said, the minutiae of detail with respect to each utterance is lost. Such an approach has the advantage of minimising the burden of discourse representation; the cognitive load associated with modelling the discourse is reduced, freeing the processor to direct effort towards other tasks, such as message generation. Clearly it would be beneficial for a speaker to allocate only minimal resources to the modelling of discourse when she is performing a cognitively demanding task.

Rather than maintain an accurate minute by minute record of the discourse, then, perhaps speakers can make some simplifying assumptions, which, in the majority of communicative situations, will hold true. Thus, for most natural dialogues where both conversation participants are co-present in time and space, it might be assumed that both can see and hear the same things, and should remember the same things about a conversation. Tracking differences in situational or textual context is unnecessary for speakers who assume there are no such differences. The speaker assumes that what is New for her is also New for her listener, while what is Given for her is, likewise, Given for her listener. In many circumstances this oversimplification will be harmless, but not always. Thus, in situations where, for example, a speaker reiterates the same message to a succession of listeners, as in Experiment Four above, the intelligibility loss associated with repetition will result in new listeners having to decode a relatively poor signal, without the

advantage of a stored representation to support recognition.

In much the same way as Lindblom seeks support for his H & H model in the characteristics of motor behaviours in general, we offer the oyster catcher (Tinbergen, 1951) as evidence, from the natural sciences, of such oversimplifying behaviour. Ethologists, such as Tinbergen and Lorenz, have studied the instinctive, species-specific patterns of animal behaviour, and have shown that some fixed-action behaviour patterns appear to be genetically pre-programmed to respond to certain stimulus features only.

Newly hatched herring gull chicks, for example, beg for food by pecking at the tips of their parents' beaks. To ascertain what stimulus characteristics are critical for eliciting this pecking behaviour, gull chicks were offered a variety of cardboard models of adult gull heads. The model they pecked at most often was long and thin and had a red patch at its tip: i.e. the model that most closely corresponded to the characteristics of an adult herring gull's beak.

A similar study was conducted on the behaviour of oyster catchers, ground-dwelling birds that nest on sand. On occasion, an oyster catcher's egg will roll off the nest, and the bird must retrieve it. By placing larger, round stones near the nest, Tinbergen (1951) showed that the retrieval behaviour was, in fact, keyed to a simplified representation of the egg: oyster catchers retrieved **the largest visible round object**. In the majority of situations this over-simplification will be harmless; there are likely to be few large round objects on the beach to challenge the egg as the most suitable candidate for retrieval. But if the nest is built in an unusually rocky environment – or an ethologist comes along and places an appropriately large object in the way – then the gap between the simplification and the true state of affairs becomes a significant one, and the consequences are potentially harmful.

Speaker behaviour which is based on the speaker's own model of the discourse, will – in most instances – not differ significantly from behaviour based on a more accurate, detailed model of what the listener knows. Where, however, the models differ quite radically, as in Experiment Four when the listener does not share the speaker's previous map experience, the simplification may be detrimental.

There is some evidence to support this claim in the results of Experiment Two. Recall that the material in this experiment involved the self-repetition of unshared landmarks, the existence of which should have been denied by the listener. A significant interaction between mention and denial was shown, by Scheffé test, to reveal a relation between denial and the intelligibility of **first** mentions: intro-

ductory mentions that failed to elicit a denial were more reduced, compared with their citation form, than introductions that were later, accurately, denied. First tokens associated with faulty replies, then, were unusually unintelligible for introductory mentions. Since we know that less intelligible tokens may be associated by the listener with Given information, it is possible that these poorly articulated introductory mentions were signalling the wrong information to the listener. In other words, the speaker's egocentric error was marking a token as accessible, and signalling to the listener to retrieve the referent from his discourse model where, it had, as yet, no representation. It may have been this error in speaker behaviour that was responsible for the failure by the listener to respond with the appropriate feedback.

It is possible, then, that the "(gross), predictive, running estimate of the implicit, listener-generated contribution [to lexical access]" that Lindblom (1983b:157) proposes speakers make is based on what the speaker herself, rather than her listener, knows. Thus the degree of hypo- or hyper-articulation associated with a token's production is more egocentrically speaker-based than might at first appear to be the case.

Chapter 7

Phonological Reduction Processes I: Place assimilation of Word-Final Nasals

7.1 Introduction

In the previous chapter it was established that the information structure of the speaker's message has a significant effect on a word's intelligibility: repeated mention of an entity already established in the discourse results in a loss of intelligibility, a phenomenon which we refer to as the repetition effect. While this appears to support Lindblom's view that speakers adjust their pronunciation of words in running speech to complement the information available to listeners in the remainder of the discourse, there is evidence to suggest that speakers' behaviour is more egocentric than is implied by the H & H theory's listener-oriented approach.

Having established that reductions in intelligibility occur with repetition, the question arises as to how such differences in clarity might be realised. It seems reasonable to suppose that the application of various phonological reduction processes prevalent in running speech (Shockey, 1974; Brown, 1977; Shockey and Bond, 1980; Dalby, 1984) may contribute in part to the loss of intelligibility evident in our data. In this chapter I assess the contribution to intelligibility loss made by place assimilation of word-final alveolar nasals in English.

Because of the difficulties in acoustic analysis of nasal segments – nasals are characterised by *lack* of formant structure and spectrographic detail – a two-pronged attack was adopted: in addition to an **acoustic** analysis based on pole-zero decomposition, a group of subjects were asked to listen to the stimuli and

make a **perceptual** judgement on the degree of evidence for assimilation. The results of the acoustic analysis were uninformative (details are summarised in section 7.4 below) so the greater part of this chapter will focus on the outcome of the perceptual study.

The perceptual study explored the relation between intelligibility (as judged by subjects listening to the excerpted token), and the degree of assimilation perceived by expert phoneticians. If assimilation contributes to intelligibility loss, then when presented in isolation, assimilated tokens should be harder to recognise than unassimilated tokens. This difference should be reflected in the intelligibility scores. We should also expect to find a repetition effect for assimilation: if second tokens are less intelligible (*vis-à-vis* citation form) than first tokens, then these second tokens should also show more signs of having undergone assimilatory processes than first mentions.

7.2 Materials

The appropriate phonological environment is, of course, a necessary pre-requisite for processes of reduction, such as assimilation, to apply. Such environments are difficult to achieve in spontaneous conversation where the experimenter has no control over the form of the utterances being produced. However, the careful selection of feature names in the HCRC Map Task Corpus (see Section 5.2.2.1) encouraged speakers to employ linguistic structures of the appropriate form. Thus some of the landmark names involved alveolar nasals in an environment which would encourage – but not enforce – the application of anticipatory place of articulation assimilation. Data was selected from the 128 unscripted conversations that comprise the HCRC Map Task Corpus (Anderson *et al.*, 1991).

Appropriate landmark names involved word-final alveolar nasals preceding both voiced and voiceless bilabial and velar oral stops. Table 7.1 provides examples of landmark names that illustrate each of the four different place/voice categories.

The dataset comprised all examples of landmark names designed to invoke nasal assimilation for which there was already a measure of intelligibility. From these were selected all cases for which there was both a first and a second mention, along with the corresponding citation form. This provided a total of 21 usable examples of nasals preceding labial stops (e.g. ***telephone*** *box*), and 13 usable examples of nasals preceding velar stops (e.g. ***lemon*** *grove*). Although a total of 14 nasal-final words preceding velars were presented in the perceptual study

Context	Landmark name	Transcription
voiceless labial	*caravan park*	[kaɾəvan paɾk]
voiced labial	*saloon bar*	[səlun baɾ]
voiceless velar	*fallen cairn*	[fɔlən keɾn]
voiced velar	*lemon grove*	[lɛmən gɹov]

Table 7.1: Examples of landmark names involving nasal assimilation in two place and two voicing contexts

described below, one item (*overgrown gully* from q4n) had to be excluded from subsequent analyses because the signal of the second token was found to have been clipped, primarily as a result of the speaker laughing.

For each of these landmarks there were two running speech tokens: the first, introductory, mention and the second mention. In some cases the second mention was by the same speaker who introduced the word, in other cases it was by a different speaker. Each running speech token had a corresponding citation form against which it could be compared. Since comparisons were exclusively between the degree of assimilation in citation form versus the assimilation in the running speech token of the same word form uttered by the same speaker, the problem of variability across speakers was less important than it might otherwise have been. In total, there were 68 running speech/citation form pairs, forming 34 sets of four tokens. Table 7.2 indicates how many of the landmarks fell into each place/voice category.

	Voicing context		
Place	Voiced	Voiceless	*totals*
Labial	18	3	*21*
Velar	5	8	*13*
totals	*23*	*11*	*34*

Table 7.2: Number of nasal-final word forms preceding oral stop consonants produced at either labial or velar places of articulation both with and without voicing

7.3 Replicating the intelligibility effect

Before assessing the contribution made by assimilation to the intelligibility loss associated with repetition, it was necessary to establish the existence of an intelligibility effect for the set of nasal word forms. The repetition effects reported in Chapter 6, while reasonably robust, were based on a large corpus of data, of

which the nasal tokens constitute a small subset. It was important, therefore, to ascertain whether the repetition effect still held for this subgroup.

For the set of word types used in this study on nasal assimilation, **mean** scores for correct recognition were calculated for all tokens which appeared in more than one intelligibility experiment. For the remaining tokens, the intelligibility score recorded for the one experiment in which the token occurred was used.

Scores for correct recognition of the nasal-final word forms were submitted to an Analysis of Variance by materials (by subjects analysis was not possible since the items were gathered from a series of different experiments involving different groups of subjects). Mean intelligibility scores for citation forms and spontaneous mentions can be found in Table 7.3.

	Speech form		
Mention	Citation	Token	***loss***
First	.697	.476	*.221*
Second	.761	.414	*.347*
mean	*.729*	*.445*	

Table 7.3: Mean intelligibility of citation forms and running speech tokens for first and second mentions of nasal-final word forms (N=34)

The ANOVA was run on **intelligibility loss**, with difference from citation as the dependent variable, mention (First/Second) as a repeated measure, and the experiment from which the data was taken (Experiment One, Two or Three: see Chapter 6) used as a between measures grouping variable. It revealed a main effect of mention: the loss of intelligibility from citation to running speech was significantly greater for second mentions than for first mentions (Mention: $F_2(1, 31) = 7.27,\ p = .01$). Given that the dataset includes other-speaker repetition, 'loss from citation' is a more informative measure of the repetition effect than a mention by form interaction of raw scores, since loss from citation controls for the variability in citation form production associated with speaker differences.

It would seem, then, that there is a significant effect of repetition on intelligibility for this subset of data, despite the relatively small sample size: second mentions of landmark names ending in alveolar nasals suffer greater intelligibility loss *vis-à-vis* their citation form than first mentions. We are now in a position to ask, therefore, to what extent the loss of intelligibility may be attributable to the place assimilation of the word-final nasal. In other words, do place assimilations behave in such a way as to make them candidate sources of unintelligibility?

To answer this question an **independent assessment of assimilation** is required for each token. If the reduction process of assimilation contributes to unintelligibility, then we would expect the level of assimilation to reflect the pattern of intelligibility loss. Specifically we would make the following prediction:

Hypothesis 7.1 *Running speech tokens will undergo more assimilation than citation forms*

Hypothesis 7.2 *Levels of assimilation will increase for second mentions of running speech tokens, compared with first mentions*

Hypothesis 7.3 *A positive correlation will be found between degree of assimilation and intelligibility loss: the more assimilated the token, the greater the loss of intelligibility*

The following sections describe two different approaches to the provision of the required assimilation score. The first was an attempt to obtain an objective measure of assimilation based on acoustic measures; the second approach was to obtain judgements of assimilation based on the auditory perception of trained phoneticians.

7.4 Acoustic study

One obvious approach to measuring assimilation is to analyse the speech data itself: to find acoustic evidence for place of articulation that is non-alveolar. Nasal segments, however, present a number of difficulties in this respect. To explain why the acoustic analysis of these nasal tokens proved problematic, I present, first, an account of the general acoustic characteristics of nasal stop consonants.

7.4.1 Acoustic characteristics of nasal stops

7.4.1.1 Frequency of formants and anti-resonances

The difficulty of establishing reliable acoustic correlates of nasalisation has long been recognised (Garvin and Ladefoged, 1963, page 197). Almost thirty years of research later, authors such as Qi and Fox (1992) are still stating that:

> *"robust features that enable automatic within-class differentiation of nasal consonants have not been well established" (1992:1718).*

The articulatory feature most characteristic of nasality is, of course, the coupling of the oral and nasal cavities: in nasal consonant production there is a closure at some point in the oral cavity combined with an open exit of air through the nasal cavity. For nasal consonants, the combined nasal-pharyngeal tract acts as the main pathway for sound, and, in terms of the standard 'Acoustic Tube Model' the oral closure is treated as a side-branching resonator. The presence of the side-branch – the blocked oral cavity – introduces an **anti-resonance**, or **zero**, in the spectrum of a nasal consonant.

The main acoustic features of nasal consonants mentioned in the literature are:

1. a prominent low frequency resonance lying somewhere between 200 and 400 Hz (House, 1957; Hattori *et al.*, 1958; Fujimura, 1962; Glass and Zue, 1986);

2. the suppression of energy in the middle of the frequency range (Glass and Zue, 1986);

3. the presence of a prominent anti-resonance (House, 1957; Hattori *et al.*, 1958; Fant, 1970).

Despite the fact that the upper formants are not always visible on spectrograms, Fujimura (1962) was able to demonstrate via analysis-by-synthesis techniques that these formants are less widely spaced than in vowels and show greater damping, which tends to lead to a general lack of spectrographic detail. In some sense, then, the *absence* of spectrographic information is itself an indicator of nasal consonant articulation.

The suppression of energy in the mid frequency range means that spectrograms provide minimal information about the place of articulation of different nasal stop consonants, except for what can be observed in the transitions into and out of neighbouring vowels (e.g. Liberman *et al.*, 1954). Although it has been demonstrated that information about place of articulation is perceptually available from the nasal murmur and not just transitions (Recasens, 1983; Kitazawa and Doshita, 1984; Kurowski and Blumstein, 1984; Repp, 1986; Kurowski and Blumstein, 1987) the acoustic bases of the capacity to make these discriminations are not immediately evident from spectrographic representations alone. Kurowski and Blumstein (1987), in attempting to find stable acoustic properties for English [m] and [n] across vowel contexts for different speakers, chose to conduct 'critical band' analyses[1], converting spectral information into a logarithmic Bark scale.

[1] chosen to reflect the properties of the human auditory system

They found significant effects in the pattern of spectral change when moving from the murmur into the release phase, which could be used successfully to distinguish labial from alveolar nasal consonants: for labials there was a greater change in energy in the region of Bark 5-7 relative to that of Bark 11-14, while for alveolars, the reverse was true. Approximately 85% of utterances could be correctly classified for place of information ([m] or [n]) by comparing the energy change in these two spectral regions.

The major distinguishing factor for place of articulation, however, appears to be the location of the primary *anti-resonance* (House, 1957; Hattori *et al.*, 1958; Kurowski and Blumstein, 1984; Repp, 1986). According to the Acoustic Tube Model, the longer the branching tube, the lower the frequency of the anti--resonance; thus labials have a low frequency zero while velars have a high frequency zero.

Table 7.4, adapted from Rooney (1990:107), summarises the findings of the key investigators in the field. As can be seen, there is tremendous variability both within place categories for any one research paper, and between the different studies. Fujimura (1962) also observes that the location of the zero (anti-resonance) for labial nasals is heavily dependent on the neighbouring vowel context. In general, labials appear to be characterised by a zero somewhere below 1200 Hz, alveolars by a somewhat higher frequency zero around 1500 Hz (but possibly as high as 3000 Hz), while velars are characterised by the absence of a zero below 3000 Hz.

Researcher	**Labial**	**Alveolar**	**Velar**
House (1957)	1000 Hz	3300 Hz	greater than 5000 Hz
Hattori *et al.* (1958)	500-1000 Hz	450-1500 Hz	"dull and complex spectrum"
Fujimura (1962)	750-1250 Hz	1600 Hz	nothing below 3100 Hz
Fant (1970)	800 Hz (3500 Hz)	1800 Hz (5600 Hz)	none
Rooney (1990)	821 Hz (3469 Hz)	1591 Hz (4029 Hz)	3939 Hz

Table 7.4: Location of anti-resonances reported by different researchers for three different nasal consonants; figures in brackets indicate location of a second anti-resonance

It is evident from the variation reported in Table 7.4 that it is difficult to locate accurately the centre frequency of the anti-resonance of a nasal consonant. Qi and Fox (1992) adopt an alternative approach, therefore, choosing to analyse the less intensive spectral perturbations closely related to the anti-resonance rather than the zero itself. They transform the spectrum using the Perceptual Linear Predictive (PLP) method, which is based on characteristics of the auditory system. Work on CV syllables, where C was either [m] or [n], revealed that the frequency and bandwidth of the second peak of the PLP spectrum was able to help distinguish labial nasals from alveolars. However, research using VC syllables has proved to be less successful, with identification results rising only slightly above chance level.

The literature also suggests that the locations of both formant (pole) and anti-formant (zero) frequencies exhibit tremendous variation between speakers, again leaving it difficult to establish normative values. Asymmetries in the nose, for example, can result in the introduction of an extra pole-zero pair between the poles of a symmetric nasal tract (Lindqvist-Gauffin and Sundberg, 1976). Indeed, Lindqvist-Gauffin and Sundberg argue that the traditional model of the nasal tract is over-simplistic, and that a more complex transfer function is required to account for the shunting effects of the two primary nasal sinuses (sinus maxillares and sinus frontales), since these are hugely variable in form from one speaker to another. More recently, Kitazawa and Doshita (1984) have found considerable effects of speaker variability on automatic nasal consonant discrimination. They conclude that:

> *"the apparent anatomical variability of the width of the nasal passages and the amount of mucus filling in cavities and constrictions are reflected in the variability of spectrographic details when data from different subjects are compared." (1984:52)*

In summary, then, it appears that finding a stable, dependable frequency value for the major formants, or indeed, anti-formants, of nasal consonants is difficult. It is especially difficult to discriminate between different places of articulation when dealing with tokens from a number of different speakers.

7.4.1.2 Murmur versus transitions

The words used in the intelligibility experiments were selected from spontaneous utterances involving landmark names which were carefully constructed to encourage the application of certain phonological processes of reduction. For this reason,

the word-final nasal segments preceded either a labial or a velar oral stop. Consequently there are no clear formant transitions out of the nasal into the following segment, since the following segment in all cases involves a stop closure.

Generally, it has been argued that the nasal **murmur**[2] serves predominantly as a cue to nasal **manner** of articulation (Pickett, 1965; Delattre, 1968; Larkey *et al.*, 1978), while the formant **transitions** into and out of the nasal murmur provide necessary and sufficient cues to **place** of articulation (Malécot, 1956; Recasens, 1983). Nevertheless, although much of the research concerning the perception of place of articulation in nasal consonants has focused on the role of formant transitions as cues to place information, it has been shown that the murmur carries place information too (Kurowski and Blumstein, 1984; Repp, 1986).

Malécot (1956) and Recasens (1983) both demonstrated that although the nasal resonance may contain a small amount of place information, it was the loss of transitions which resulted in a significant decrease in identifiability of place of articulation. However, since their work involved the juxtaposition of murmurs with inappropriate transitions (for example, a labial murmur followed by a vowel transition out of an alveolar) it is possible that artificial spectral discontinuities may have contributed to the perceptual salience of the transition. Kurowski and Blumstein (1984) avoided this potential pitfall by using only natural speech tokens, combining each of two nasal consonants [m, n] with each of five vowels [i, e, a, o, u]. While they found no systematic difference in duration, amplitude, or fundamental frequency across place of articulation, they did find that:

> *"the murmur seems to provide a fairly reliable cue to place of articulation in edited nasal consonants, and, moreover, it is about as effective a cue as the formant transitions" (1984:387).*

Repp (1986) essentially confirmed the findings of Kurowski and Blumstein extending the analysis to cover tokens from more than a single speaker[3].

So it would appear, then, that the murmur – or resonance – may contain information which can aid listeners in identifying the place of articulation of the nasal consonant. Thus, it is not impossible to find evidence of acoustic differences between nasal consonants made at varying places of articulation, even in the context

[2]the murmur may be defined as "the sound produced with a complete closure at a point in the oral cavity, and with an appreciable amount of coupling of the nasal passages to the vocal tract" (Fujimura, 1962, page 1865)

[3]In both studies significant effects of vowel context were found; in particular Repp found that for nasal consonants preceding the high front vowel [i] accurate place identification was only possible with the simultaneous presence of both murmur and transition components.

of a following oral stop consonant when transitions out of the nasal will not be present.

7.4.2 Pole/zero decomposition

The literature predicts that the location of the primary anti-resonance – or zero – is one of the few ways to distinguish place of articulation of nasal stop consonants: while labials are characterised by a zero-resonance in the region of 1200 Hz, velars are associated with an absence of a zero below 3000 Hz. The anti-resonance in alveolar articulations is located somewhere between the two, usually around 1500 Hz but sometimes as high as 3000 Hz (see Table 7.4).

The accurate analysis of zero-resonances has presented problems for traditional methods of acoustic analysis such as spectrography and linear prediction, which generally *ignore* the pattern of zeros. The group-delay spectrum (Broe, 1993), based on the derivative of the phase rather than magnitude spectrum[4], has been shown (Yegnanarayana, 1981) to yield superior resolution and discrimination of both poles (resonances) and zeros (anti-resonances).

Rooney (1990) demonstrated that pole-zero decomposition based on Yegnanarayana's technique can be used effectively to identify different speakers' productions of the nasal consonant [ŋ] in a verification task. Success in distinguishing nasal productions requires accurate location of zero-resonances: the group-delay spectrum is characterised by a sudden excursion from a zero base-line (Broe, 1993) which facilitates the detection of zeros in the frequency response.

7.4.2.1 Procedure

Each word token used in the intelligibility experiments was associated with its own sampled speech datafile. The files corresponding to the set of nasal-final words were segmented at a phonemic level, with segmentation lines drawn between each phonemic segment of the nasal word, following the guidelines detailed in Chapter 5. The same ESPS XWAVES software was used as had been employed in locating word boundaries for all intelligibility stimuli. Both spectrograms and time/amplitude waveforms were used to supplement auditory information.

Once the nasal segment had been located the group-delay spectrum was deter-

[4]The group-delay response represents the *rate-of-change* (with respect to frequency) of the phase response to a particular pole/zero. The rate of change increases when passing through frequencies in the vicinity of a pole (or zero).

mined, using a program based on Yegnanarayana (1981) and written by Dr M. Broe to run on ESPS-headered speech files in a UNIX environment. The program prompts for input figures relating to:

- the range (amount of speech) over which the program should run;
- the number of cepstral coefficients to be used[5];
- the number of FFTs to be performed;
- the number of poles to plot;
- the number of zeros to plot.

Input values are determined by trial, although it is recommended to plot twice as many poles as zeros (Rooney, 1990). Unfortunately the plots produced by the program are non-manipulable images; the frequency of the zero is most easily determined by eye, or measured by hand from a screen dump[6].

7.4.2.2 Results

The first set of analyses was run over the initial 200 samples (12.5 ms) of the nasal segment, using 20 cepstral coefficients, 9 FFTs, 5 poles and 5 zeros. The plots produced were extremely variable with no obvious pattern to the location of zeros, even for citation form productions with clear canonical-sounding [n] segments.

In the second trial the input values were changed to 25 cepstral coefficients, 9 FFTs, 25 poles and 12 zeros. The results were no more consistent than before.

The third trial followed recommendations from Rooney (personal communication), with input values set to 25 cepstral coefficients, 9 FFTs, 10 poles and 5 zeros, the frame length set to 300 samples (18.75 ms), and with poles and zeros measured at three locations within the nasal:

start – the first 300 samples from the onset of the nasal segment;

mid – 150 samples either side of the nasal midpoint;

end – the last 300 samples up to the offset of the nasal.

[5]The number of cepstral coefficients determines the smoothness of the plotted curve.

[6]Alternative methods of plotting were explored but proved problematic due to the vagaries of the XWAVES software.

The resulting plots were still erratic, with the frequency location of zeros for canonical [n] segments in citation form varying not just for different tokens but for different time locations (start/mid/end) within the same token. Frequency values for zeros ranged from less than 1000 Hz to over 4000 Hz.

Given that the spectral characteristics of nasal segments are known to exhibit tremendous variation across speakers (Lindqvist-Gauffin and Sundberg, 1976; Kitazawa and Doshita, 1984; Rooney, 1990) it was decided to run a control with a single speaker uttering nasals of known place of articulation, in order to establish some base-lines. I therefore recorded myself articulating citation form productions of landmark names which were:

1. unassimilated, lexically alveolar
   e.g. [kaɹəvan pɑk] (*caravan park*), [fɔlən kɛən] (*fallen cairn*)

2. assimilated, lexically alveolar
   e.g. [kaɹəvam pɑk] (*caravan park*), [fɔləŋ kɛən] (*fallen cairn*)

3. unassimilated, lexically labial/velar
   e.g. [θim pɑk] (*theme park*), [fɔlɪŋ kɛən] (*falling cairn*)

Figures 7.1-7.6 represent frequency plots of zero resonances at the start, mid and end of each nasal for canonical, assimilated and lexically labial/velar nasal segments from the landmarks *caravan park* and *fallen cairn*. It is evident from the figures that the variation within each single token – let alone between tokens of the same segment by the same speaker – is such that there is no likelihood of finding a consistent pattern of zero location that could be used to discriminate place of articulation.

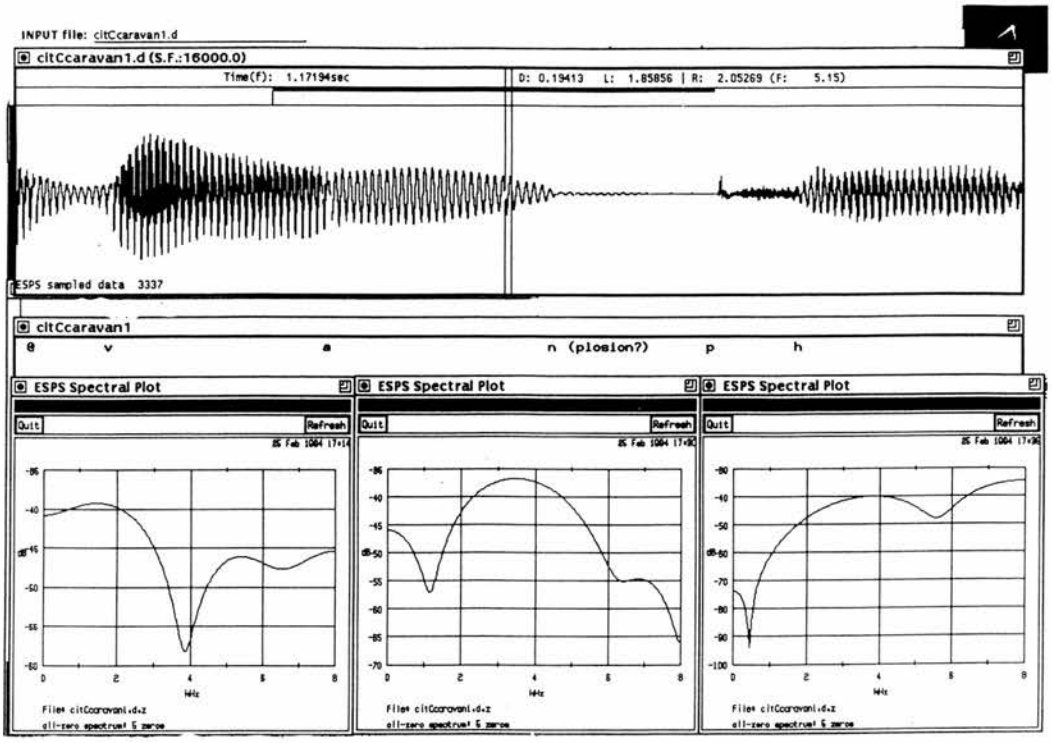


Figure 7.1: location of zero in *caravan park*: [kaɹəvan pɑk]

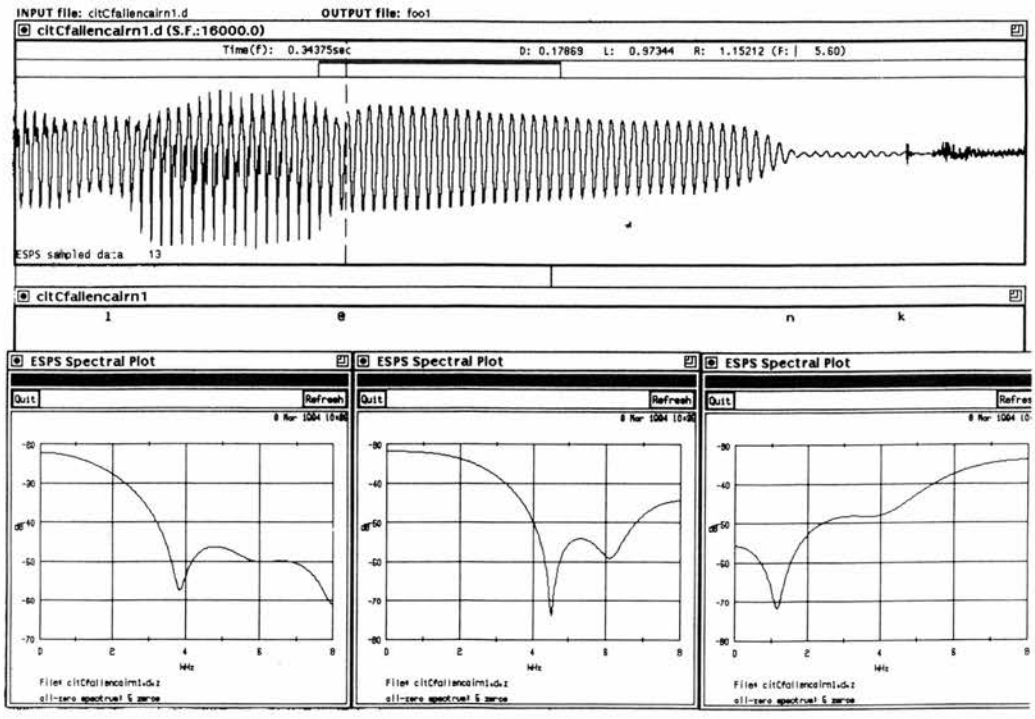


Figure 7.2: location of zero in *fallen cairn*: [fɔlən kɛən]

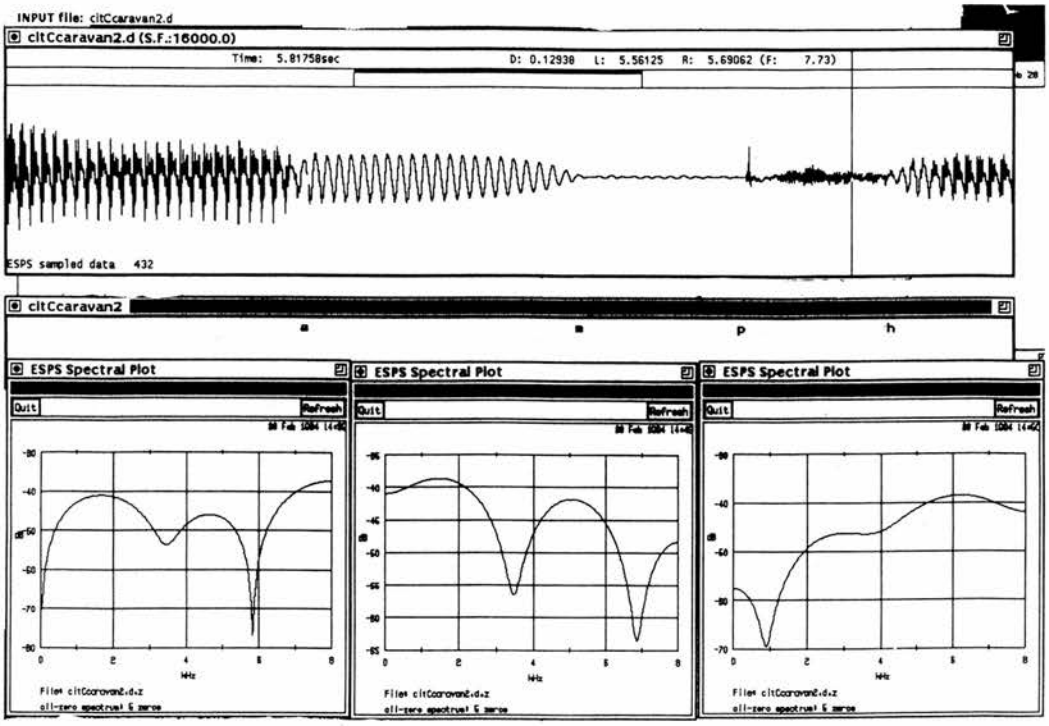


Figure 7.3: location of zero in *caravan park*: [kaɹəvam pɑk]

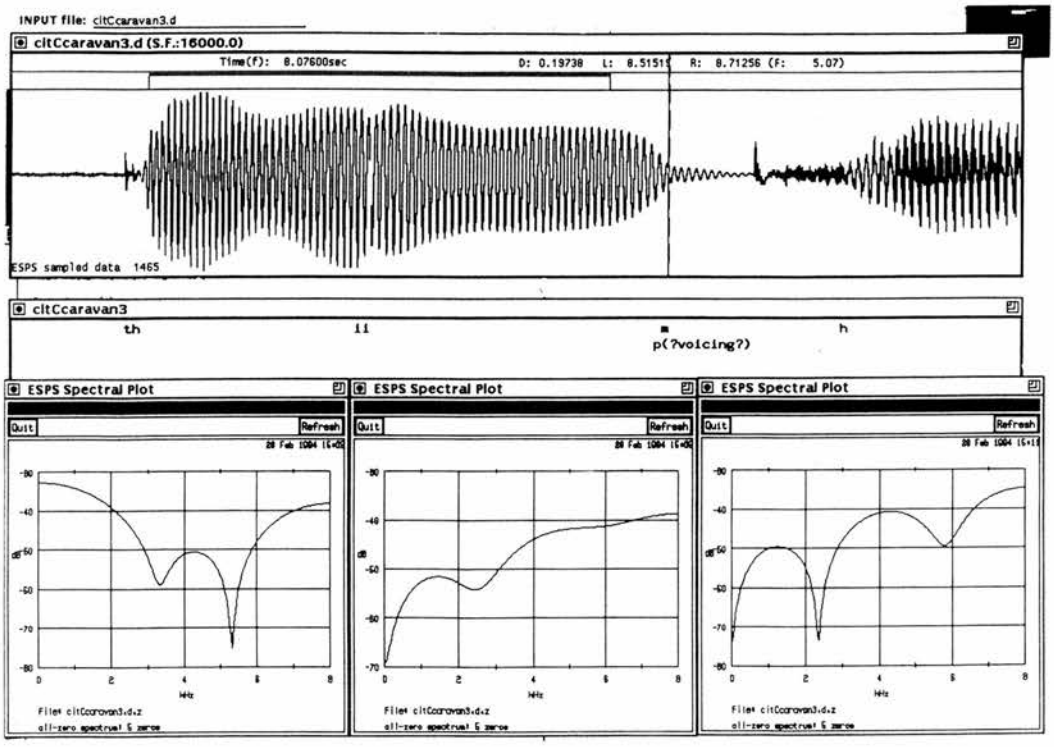


Figure 7.4: location of zero in *theme park*: [θim pɑk]

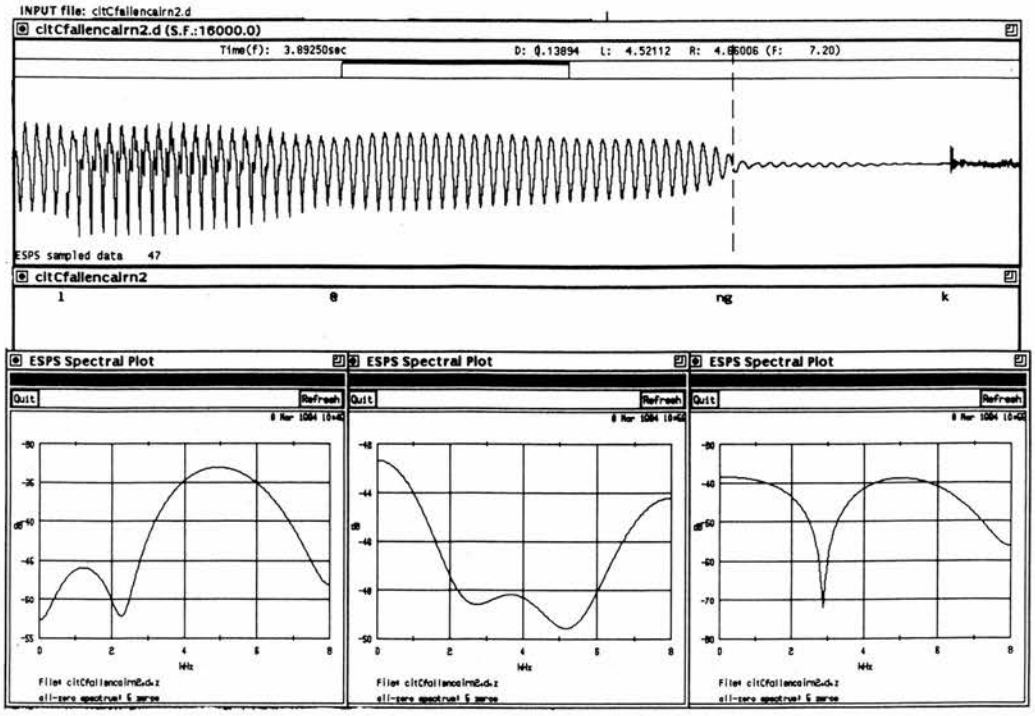


Figure 7.5: location of zero in *fallen cairn*: [fɔlən kɛən]

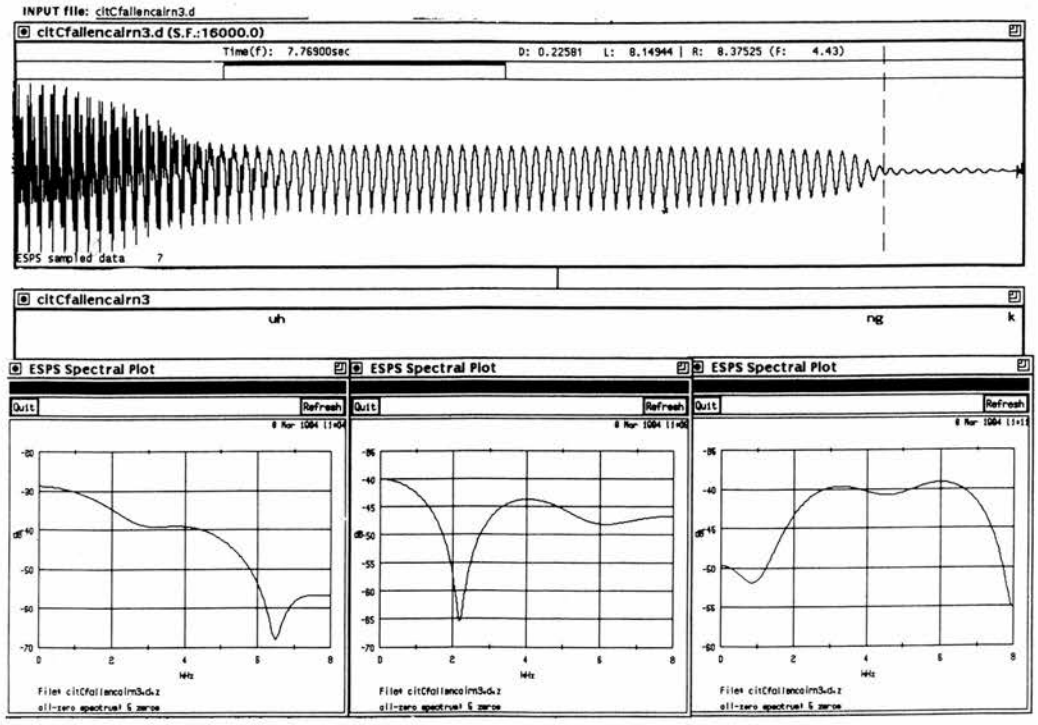


Figure 7.6: location of zero in *falling cairn*: [fɔlɪŋ kɛən]

7.4.2.3 Conclusion

The analysis of the nasal assimilation data in terms of pole-zero decomposition (Yegnanarayana, 1981) failed to elicit any stable findings. It proved to be impossible to establish any reliable measures of anti-resonance frequency for even carefully articulated nasal segment 'controls' recorded by the author. Given the additional variance introduced by using tokens from a large pool of different speakers, the chance of finding anything like a consistent frequency value for the various nasal tokens was minimal. It was necessary to conclude that this approach was not going to achieve any worthwhile results, and it was abandoned.

Why the current materials were unsuitable for pole-zero analysis is of interest. Certainly they differ from the materials usually employed: multiple tokens from just a few speakers produced in carefully controlled contexts, usually achieved by asking subjects to read aloud specially designed phrases. The selection criteria for appropriate tokens in particular and the design of the HCRC Map Task Corpus in general gave us few tokens spoken by any one speaker, with speaker variability consequently adding to the general noise. Since these speakers were also linguistically naive, asking them to produce clearly unassimilated versus completely assimilated tokens during the list reading sessions, for example, was not an option.

Further, the attempt to induce assimilation by placing word-final nasals before labial and velar stop consonants meant that potentially valuable cues to place of articulation in the transitions out of the nasal were lost.

7.5 Perceptual study

7.5.1 Introduction

Given the failure to achieve a satisfactory, informative result using objective acoustic parameters, possible **perceptual** effects of assimilation were sought by presenting word tokens to groups of experienced subjects and asking them to assess the degree of assimilation they heard for each token. Although this approach results in a subjective assessment of assimilation, it does in fact reflect the nature of the intelligibility responses, which are themselves based on perceptual judgements. The difference is that in the former case, subjects are asked to make judgements on the quality of nasal production, independent of lexical information about the word itself, while in the latter task, subjects are asked to make a

decision about lexical identity.

7.5.2 Method

The perceptual study as run had two components:

- Part A = single words excised from context, e.g. *saloon, lion*
- Part B = whole landmark names, e.g. *saloon bar, lion country*

Subjects were told to complete Part A before going on to Part B. Because the stimuli for the intelligibility studies were all presented as single word tokens, excerpted from all context, it was the judgements of assimilation for words heard under this condition (i.e. Part A) that were required in order to know what information was or was not available to the original intelligibility subjects in attempting to recognise the word. For this reason I present below only the results pertaining to Part A of the exercise[7].

The questions to be addressed are:

- Are experts' judgements in line with our phonological expectations? That is, do experts give a high [m] rating for nasals preceding labial stops and a high [ŋ] rating for nasals preceding velar stops?
- Are there any consistent differences between first mentions, second mentions and citation forms in line with the intelligibility results? For example, do we find more cases of assimilation for tokens than citation forms, and more assimilation for second mentions than for first mentions?
- In general, do the assimilation judgements correspond to the intelligibility results? Are assimilated tokens less intelligible than unassimilated tokens?

A satisfactory answer to the first question is a necessary prerequisite for asking the rest.

[7]The main result from analyses of Part B was that the experts were subject to much the same contextual effects as listeners with non-trained ears: subsequent context significantly reduced the judgements of inappropriate assimilation (judgements of [m]-ness for nasals preceding velars and of [ŋ]-ness for nasals preceding labials) to virtually zero. Judgements of [n]-ness remained unchanged. The presence of the following word appears to block the percept of assimilation to the inappropriate place of articulation.

7.5.2.1 Materials

A total of 118 speech stimuli were presented to each subject. Each stimulus was a single word token which ended in a lexical alveolar but which might have undergone assimilation. The 118 stimuli comprised multiple tokens of 35 separate items (see Section 7.2 above). Every item was associated with two running speech tokens – the first and second mentions of the landmark name in the discourse – and either one or two citation form tokens: one citation form for same-speaker repetitions (N=22), but two citation forms for other-speaker repetitions (N=13). This resulted in $(22 \times 3) + (13 \times 4) = 118$ individual tokens.

7.5.2.2 Procedure

Since all tokens had been used previously for the intelligibility experiments they were already digitised with start and end points labelled (see Section 5.3 for details of segmentation criteria). While for the experiments on intelligibility tokens had been overlaid with noise, the tokens presented to subjects in the perceptual task were not masked: it was anticipated that the task of identifying place information from the nasal segment would be difficult enough without contending with the degradation introduced by noise overlay.

The order of presentation of the stimuli was randomised using the last 4 digits of the duration value. Since duration was stored as a 7 digit figure (e.g. the duration of *saloon* taken from the first mention in dialogue q1ec3 was 231.1937 ms) the randomisation bore no relation to the actual durations of the stimulus words. Each stimulus was preceded by a numerical identifier and was presented twice, with an ISI between repetitions of 30 ms, and an ISI between different stimuli of 3 seconds. A total of 30 different speakers were represented, both male and female. Stimuli were recorded onto standard audio cassette tapes, and given to experts together with a response booklet. Subjects completed the task in their own time, having been recommended to select a quiet environment and to complete the task within a single session if possible. Subjects were permitted to listen to the stimuli as many times as they felt was necessary to make a judgement.

7.5.2.3 Task

Subjects were asked to rate each word-final nasal on three scales: labial, alveolar, and velar. A rating of 0 indicated that no evidence was perceived to suggest the consonant was produced at this place, while a rating of 5 indicated that the

perceptual evidence was fully consistent with an articulation at this place. The three options were not mutually exclusive, so that in principle it was possible to assign a rating of 5 on more than one scale for any one token. An example extract from a response sheet is presented in Figure 7.7.

8: saloon

[m]	0	1	2	3	4	5
[n]	0	1	2	3	4	5
[ŋ]	0	1	2	3	4	5

Comments:.....

Figure 7.7: Example section of response sheet for the stimulus word *saloon*

7.5.2.4 Subjects

Four of the experts (e1-e4) were professional phoneticians or phonologists at the University of Edinburgh. The remaining five phoneticians (e5-e9) came from Cambridge University. The results presented here collapse the two sources of data into one set unless stated otherwise.

7.5.3 Results

7.5.3.1 Were tokens perceived as assimilated?

Let us consider, first, whether the perceptual experiment was an appropriate means of eliciting scores of assimilation. Were experts able to judge any of the tokens as assimilated, and were their judgements in the appropriate direction, i.e. judged as sounding [m]-like preceding a labial context, and [ŋ]-like preceding a velar context?

Each stimulus received a three-way judgement of assimilation from each expert subject: a judgement of [m]-ness (from 0-5), a judgement of [n]-ness (from 0-5) and a judgement of [ŋ]-ness (from 0-5). The set of 3 place values will be referred to as a **triplet** judgement. Each triplet judgement (e.g. '0-4-1', where the first value is for [m]-ness, the second for [n]-ness, and the third for [ŋ]-ness) was classified according to whether:

- there was a clear alveolar percept with **no assimilation**, e.g. 0-5-0, 0-4-0, 0-3-0

- there was a **dominant labial** judgement, e.g. 5-0-0, 5-2-3, 3-2-0
- the **labial** judgement was **greater** than the **velar** (while less than or equal to the alveolar judgement), e.g. 3-3-0, 2-3-1, 1-5-0
- there was a **dominant velar** judgement, e.g. 1-0-4, 0-0-5, 1-2-3
- the **velar** judgement was **greater** than the **labial** (while less than or equal to the alveolar judgement), e.g. 0-3-2, 1-4-2, 1-3-3
- the labial and velar judgements were the **same** and greater than zero, e.g. 3-0-3, 2-3-2, 1-4-1

This categorisation can be seen to distinguish cases of clearly unassimilated tokens (those classed as clear alveolar percepts with no [m] or [ŋ]-ness heard at all) from tokens which are perceived to involve some degree of assimilation. The number of triplet judgements that fell into each of the above categories was summed for all 9 experts. Judgements of [m]-ness for words preceding labials, and of [ŋ]-ness for words preceding velars, were classed as examples of (some degree of) assimilation in the **expected** direction; [m] judgements for words preceding velars, and [ŋ] judgements for words preceding labials were classed as examples of assimilation in the **opposite** direction. The results, presented as percentages of the total number of judgements made, can be seen in Table 7.5, where 'exp' stands for 'expected', and 'opp' stands for 'opposite'.

	Following context		
	Labial	**Velar**	
Percept	**(N=84)**	**(N=52)**	*Mean*
alveolar	34.40	33.33	*33.98*
expected	11.57	24.49	*16.51*
exp > opp	28.92	29.80	*29.26*
exp < opp	17.05	8.08	*13.62*
opposite	5.78	2.53	*4.54*
exp = opp	2.28	1.77	*2.09*
TOTALS	100.00	100.00	*100.00*

Table 7.5: Percentage of total judgements in each percept category for nasals preceding labials and velars

One third of all tokens (33.98%) were judged as clearly alveolar, i.e. unassimilated. Around one sixth of the judgements (16.51%) showed a strong percept in the expected direction, with half as many dominant labial percepts (11.57%) as velar

(24.49%). Over one quarter of all judgements (29.26%) were perceived as primarily alveolar, but with some evidence of appropriate assimilation ('exp > opp' in Table 7.5). It thus appears as if experts did indeed perceive some assimilation in the tokens with which they were presented.

If we examine the triplet judgements and graph the data in terms of **dominant** versus **weak** percepts – where a dominant percept is a triplet judgement with a strongly expressed preference for nasal production at one particular place of articulation (i.e. categories 'alveolar', 'expected' and 'opposite' in Table 7.5 above), and a weak percept is a triplet judgement which is predominantly [n] but with some non-canonical 'colouring' of either [m] or [ŋ] (i.e. categories 'exp > opp', 'exp < opp' and 'exp = opp' in Table 7.5 above) – then we find, first, that pre-labial nasals were perceived as more weakly assimilated than pre-velars, with few dominant [m] percepts, and secondly, that pre-labial nasals elicited more percepts of inappropriate assimilation (i.e. judgements of [ŋ]-ness) than did pre-velars (judgements of [m]-ness) (see Figure 7.8).

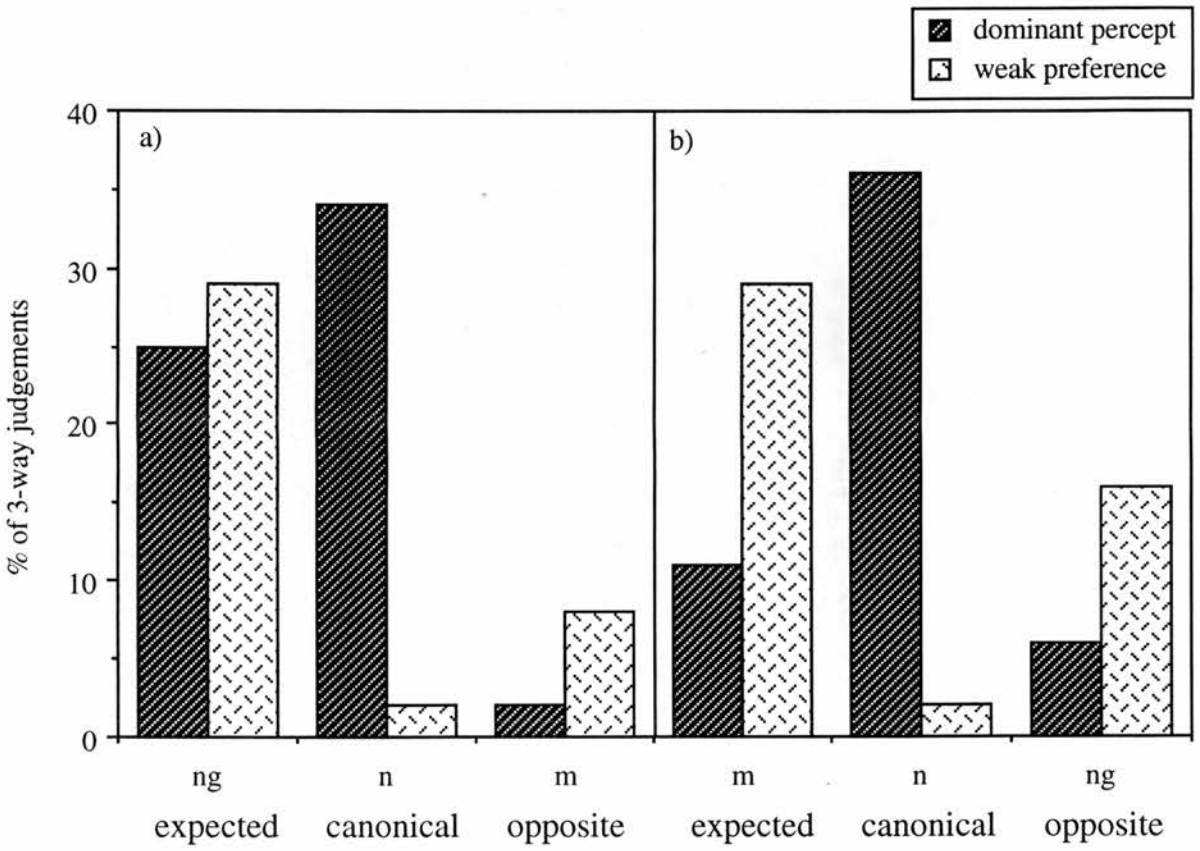
7.5.3.2 Form and mention effects

A series of 2x2 ANOVAs was run on *mean* place judgements, to explore the effects of form and mention on assimilation scores. Separate analyses were conducted for words preceding labial and velar contexts, with form (Citation/running speech Token) and mention (First/Second) as repeated measures, and voicing of the following segment as a grouping factor between items[8].

Citation forms were found to have significantly higher [n]-ness scores than running speech tokens for words preceding both velar and labial stops (pre-labials: $MinF'(1,26) = 4.39, p < .05$; pre-velars: $MinF'(1,15) = 7.36, p < .025$). Citation forms were also perceived as sounding less assimilated than the corresponding tokens excerpted from natural dialogue. Judgements of *appropriate* assimilation ([m] pre-labial, [ŋ] pre-velar) were significant by subjects and approached significance by materials for the pre-velar nasals (pre-labials: $F_1(1,8) = 6.59, p < .05$; $F_2 < 1, n.s.$; pre-velars: $F_1(1,8) = 26.74, p < .001$; $F_2(1,11) = 3.48, p = .0891$). There are also significant effects for *inappropriate* judgements of assimilation, with tokens preceding velars receiving significantly higher [m]-ness scores than citation forms (pre-velar: $MinF'(1,19) = 4.70, p < .05$) and tokens in labial

[8]The only significant effect of voicing to be found was for inappropriate judgements of [ŋ]-ness preceding labials, where [ŋ]-ness scores were significantly higher for words preceding [b] than for [p] (pre-labials: $F_2(1,19) = 4.93, p < .05$). There were no significant interactions with voicing and either form, or mention, or both.

Figure 7.8: Experts' judgements of assimilation for nasals preceding
a) velar stops
b) labial stops



contexts judged as sounding more [ŋ]-like than their matched citation forms[9] (pre-labial: $F_1(1,8) = 6.22, p < .05$; $F_2(1,19) = 2.72, p = .1$).

Thus, citation forms tend to be perceived as more canonically [n]-like and less [m]- or [ŋ]-like than their corresponding running speech tokens, and this trend is stronger for words preceding velars than it is for words preceding labials. Citation forms are less likely to receive judgements of assimilation, whereas tokens may well be judged as sounding 'non [n]-like' in either or both directions (i.e. in the direction of either appropriate or inappropriate assimilation), as if experts can hear something 'qualitatively odd' about these tokens but are unclear as to the source of the oddness. (Recall that the experts heard these isolated words without any indication of the following context.)

As Table 7.6 shows, there is an effect of repetition – similar to that observed for intelligibility – found for judgements of [ŋ]-ness with second mentions judged as more assimilated (*vis-à-vis* citation) than first mentions, which do not differ from citation form (pre-velar: Form x Mention: $F_1(1,8) = 6.89, p < .05$; $F_2(1,11) = 4.55, p = .0562$; Scheffé test: t2 > c2; t1 = c1 = c2; $p < .01$). This repetition effect is significant by subjects for both appropriate and inappropriate assimilation but not significant by materials for velars preceding labials (pre-labial: Form x Mention: $F_1(1,8) = 6.22, p < .05$; $F_2 < 1, n.s.$; Scheffé test (by subjects): t2 > c2; t1 = c1 = c2; $p < .01$).

a) Preceding Labials

Mention	Citation	Token	*loss*
First	1.01	1.49	*.48*
Second	1.22	1.32	*.10*
Mean	*1.12*	*1.41*	

b) Preceding Velars

Mention	Citation	Token	*loss*
First	1.64	1.93	*.29*
Second	1.44	2.13	*.69*
Mean	*1.54*	*2.03*	

Table 7.6: Mean expert judgements of appropriate assimilation (m-ness pre-labial, [ŋ]-ness pre-velar) for citation forms and running speech tokens for first and second mentions, in words preceding labial and velar stop consonants

There is no corresponding effect for judgements of m-ness. It may be that this was due to the generally lower scores for [m]-ness overall: experts appeared to

[9] significant by subjects but not by materials

be more reluctant to judge tokens as sounding strongly [m]-like than they were to judge tokens as sounding strongly [ŋ]-like. Such results may reflect a 'lowered ceiling' effect with judgements for [m]-ness not being strong enough to reveal any insightful differences between the variables of concern.

In conclusion, there are significant form effects, with citation forms sounding more [n]-like, and running speech tokens tending to sound more [m]- and [ŋ]-like. There is also a significant repetition effect for judgements of [ŋ]-ness: second tokens are more likely to be judged as sounding assimilated than introductory mentions. It appears there is an asymmetry in the assimilation judgements, with clearer patterns emerging for the subset of data which precedes velar stop consonants, than for the subset which precedes labials.

7.5.3.3 Assimilation and intelligibility

We have now established that experts were able to perceive some degree of assimilation, and that running speech tokens tended to be perceived as less [n]-like and more assimilated than their corresponding citation forms, when heard in isolation. We have also replicated the repetition effect (second tokens are more assimilated *vis-à-vis* citation form, than first mentions) for one half of the data: the words which precede velars.

A series of correlations shows, however, that although assimilation is related to intelligibility it does not account for all of the intelligibility differences in the data. Again, there is an asymmetry in the results: pre-labial and pre-velar nasals behave differently.

Significant correlations between intelligibility and mean place judgements were found only for words preceding velar stops: the more [ŋ]-like (i.e. assimilated) words were significantly *less* intelligible ($N = 52, r = -.409, p < .005$), while the more [n]-like (unassimilated) words were *more* intelligible ($N = 52, r = .491, p < .001$). In addition, non-assimilatory non-target pronunciation ([m]-like character in a velar context) was also found to correspond with decreased clarity ($N = 52, r = -.361, p < .01$)[10].

For words preceding labial stops, however, analogous correlations were not significant ($N = 84, r = -.105, n.s.$; $N = 84, r = .184, n.s.$; and $N = 84, r = -.177, n.s.$, for expected assimilation, canonical [n] judgements, and inappropri-

[10]This result is not so surprising if we treat a phonetician's [m] response to a pre-velar nasal as a signal that there is something atypical, or 'odd' about the token's production. This same oddness may be responsible for eliciting incorrect responses from the intelligibility subjects.

ate assimilation respectively).

It would appear, for half of the data at least, that assimilation does correspond to intelligibility when words are presented in isolation: unassimilated tokens, that is, those with clearly articulated word-final [n] segments, are significantly easier to recognise. The reduction in intelligibility found when experts hear some degree of assimilation (at least for words preceding velars) introduces the possibility that other word candidates (those ending in [m] or [ŋ]) may have been activated, and that these may be competing for recognition with the target item. If a poorly articulated token of *lemon*, for example, in the context of the landmark name *lemon grove*, sounds more [ŋ]- than [n]-like, it is possible that alternative lexical items such as *lemming* will increase the competition resulting in failure to recognise *lemon* successfully.

As in the analyses of form and mention effects on place judgements described in the previous section, significant effects for words preceding velars are not replicated in the set of words preceding labials. Since the word forms in the pre-labial and pre-velar datasets had not been matched across conditions, it was possible that the observed asymmetry in results was a consequence of the pre-labial set of words containing more word-final [n] segments in reduced, or metrically Weak, syllables, which might have made the segment more difficult to perceive. When each word form was classified according to whether the word-final nasal was in a metrically Weak or Strong syllable (e.g. *seven* vs. *saloon*) a roughly equal distribution of each syllable structure was found for the pre-labial and pre-velar word sets (pre-labial: Weak=10, Strong=11; pre-velar: Weak=5, Strong=8). When the ANOVAs were re-run with metrical structure of the word-final syllable (Weak/Strong) as a between items grouping factor, the main result was an effect on [n]-ness scores, with the strength of [n]-ness judgements decreasing in reduced syllables (MetStr: $F_2(1, 32) = 6.95, p = .01$). While the means for judgements of assimilation appear to increase for reduced syllables the difference is not significant (MetStr: $F_2(1, 32) = 1.2, n.s.$). No relation was found between metrical structure and the place of articulation of the following stop consonant: the asymmetry in assimilation judgements for words preceding velars and those preceding labials cannot be explained in terms of a difference between Weak and Strong syllables (Place x MetStr: $F_2 < 1, n.s.$).

In the next section, therefore, two alternative hypotheses are offered to account for the observed asymmetry.

7.5.3.4 Why do velar and labial contexts differ?

In the analyses of the nasal assimilation data described in the sections above the results depend on whether words precede a velar or labial segment. There are at least two possible sources of explanation for this asymmetry: the first has a phonetic basis, the other a lexical one.

One hypothesis is that there is a **phonetic** difference between the assimilatory processes for pre-velar and pre-labial nasals: in other words, assimilation from alveolar to labial is qualitatively different from assimilation from alveolar to velar. The source of such a difference could be articulatory and/or acoustic in nature.

In terms of articulatory phonology (Browman and Goldstein, 1989, 1990) both labial and velar assimilation involve the temporal overlap of **distinct articulatory gestures** on separate articulatory tiers. In the case of labial assimilation the production of lip closure is entirely separate and independent from that of tongue tip contact with the alveolar ridge. Thus it is quite possible for nasals in this category to start out as alveolar, but sound labial by the time the velum is raised again, as a consequence of early lip closure. Alveolar to velar assimilation involves a combination of tongue tip and tongue body gestures. Although tongue tip and body clearly belong to the same physiological organ there is plenty of evidence to suggest that these gestures can be treated as independent from each other (see, for example, Nolan, 1992; Ellis and Hardcastle, 1997).

It seems more likely, therefore, that any kind of phonetic basis for the difference in the way velar and labial assimilations are perceived will lie in the varying **acoustic** properties of these segments. As Ladefoged (1982) observes,

> *"The difference between each of the nasals is most plainly marked by the different formant transitions that occur at the end of each vowel. There is a clear downward movement of the second and third formants before the bilabial nasal [...] and a clear coming together of the second and third formants before the velar nasal" (1982:184).*

This characteristic "coming together" of the second and third formants is sometimes referred to as the **velar pinch** and is well documented. Those lexical alveolar tokens preceding velar contexts which exhibited such a pinch would be likely to cue a strong perception of [ŋ]-ness. Unfortunately, given the lexical alveolar status of these tokens, and the likelihood of assimilation occurring as a result of gestural overlap rather than blending, it is probable that the onset of the nasal will rarely show the distinctively velar pinch described by Ladefoged above.

Indeed, an investigative examination of the tokens used in this experiment suggests that, rather, the velar quality is likely to increase through the nasal, when formant structure is much weaker and more difficult to assess.

Recent literature provides evidence to suggest that the spectral characteristics of velar **bursts** are generally more effective at cueing place of articulation than bursts made at other places of articulation (Smits *et al.*, 1996; Green and Norrix, 1997). While the burst of the conditioning context was excluded from the stimuli presented to the subjects in the perceptual experiment described in this chapter and could not therefore contribute to the judgements of [m]- and [ŋ]-ness directly, it is interesting to note that the asymmetry observed in this study is reflected in similar sorts of work reported elsewhere. It suggests, amongst other things, that there may be acoustic cues to velar identity (formant pinch, release burst) which are more robust or perceptually salient than their equivalent cues to labiality. This would go some way to explaining the difference in mean [ŋ]-ness and [m]-ness scores. The acoustic distinctiveness of velar formant transitions is likely to boost auditory judgements of assimilation for words preceding velars.

Given the observation that certain kinds of assimilation (e.g. velar) may be easier to perceive than others (e.g. labial) it should be noted that the phoneticians' perceptual judgements of assimilation will not be an absolutely reliable guide to the comparative level of assimilation for different places of articulation; in other words, a score of '2' for labial assimilation may not reflect the *same degree* of assimilation as a score of '2' given for a pre-velar token.

The second explanation is based on **lexical** differences: perhaps the distribution of alternative candidates in the mental lexicon reflects a similar imbalance between word-final labial and velar nasals? If there are more competing candidates ending in [m], for example, then these might be reducing the intelligibility of our words simply because there are many more candidates from which to choose (making recognition harder, therefore). If, in addition, these alternative competitors occur more frequently in English than our nasal assimilation word, then a poor correlation between perceived assimilation and intelligibility may simply reflect the fact that the target was more difficult to recognise because of its competitor set, independent of whether or not it sounded [m]-like. To explore whether this was indeed the case, the responses of the original intelligibility subjects were examined to see what evidence they provided in support of this hypothesis. These results are discussed briefly in the following section. A more detailed treatment is presented in Chapter 10.

7.5.3.5 Incorrect intelligibility responses: a preliminary investigation

In Section 7.5.3.3 above we found that words preceding velar stops were less intelligible when their nasals were more assimilated. To learn why pre-labial words did not show the same effect, the alternative responses of the original subjects in the intelligibility studies were analysed, to see whether the asymmetry could be explained in terms of different sets of competing lexical items.

All responses were classed according to their word-final segment ([m], [n], or [ŋ]), and then compared with the responses of the experts, by running a series of correlations. Words judged by experts as [m]-like elicited more incorrect identifications ending in [m]. This was true both of assimilated tokens preceding labials ($N = 84, r = .212, p = .05$) and of tokens preceding velars which were judged by experts as sounding (inappropriately) [m]-like ($N = 52, r = .313, p = .02$). In effect, our intelligibility subjects seemed to share an impression of labial quality with our expert judges.

Judgements of [ŋ]-ness, however, did not uniformly correlate with [ŋ]-final intelligibility responses. Words preceding velars showed the effect ($N = 52, r = .418, p = .002$), but words preceding labials and judged (inappropriately) as sounding [ŋ]-like did not ($N = 84, r = -.076, n.s.$). For judgements of [ŋ]-ness, then, experts and intelligibility subjects did not always respond to the same acoustic cues.

Perhaps intelligibility subjects were responding to nasal-final targets with -ING-ending words, independently of how the word-final nasal sounded? A closer inspection of the data, however, revealed that some of the pre-labial words which experts heard as [ŋ]-like – and which ought, therefore, to have supported an -ING-biased response – *failed* to elicit [ŋ]-ending responses from intelligibility subjects. Furthermore, very few words elicited [ŋ]-ending responses from the intelligibility subjects when they had been judged by phoneticians as not sounding noticeably [ŋ]-like. Thus, it does *not* appear to be the case that intelligibility subjects were biased – by the presence of multiple word candidates ending in the productive -ING affix in English – into responding with [ŋ]-ending words (such as *roaming*) to pre-labial targets (like *Roman*[11]) irrespective of nasal quality.

[11] excerpted from *Roman baths*

7.6 Discussion

The acoustic analysis of nasal segments in terms of pole-zero decomposition was unable to provide stable, reliable values and was consequently abandoned. The results of the perceptual experiment can be summarised as follows:

- Running speech tokens tend to undergo more assimilation than citation forms, which are more canonically [n]-like;
- There's a place-dependent relationship between assimilation and intelligibility:
  - pre-velar, but not pre-labial, nasals show a repetition effect, with more assimilation in second tokens;
  - pre-velar, but not pre-labial, nasals show a negative correlation between intelligibility and assimilation.

While the process of assimilation is able to explain certain differences in intelligibility, it is clear that there are additional factors which influence the ease with which a word is recognised, including the phonetic context in which a word appears: there is an asymmetry between responses to words preceding labials and words preceding velars.

Thus although there *is* a relation between intelligibility and assimilation – tokens of a word which are perceived to have been assimilated result in poorer recognition when excerpted from context – it is complicated by the observed asymmetry. The failure of perceived assimilation to account for the repetition effect for the subset of data preceding labial contexts indicates that there are other factors at play.

Notwithstanding this asymmetry in results, it is evident that there is a cost involved in the processing of assimilated tokens. Without supporting context, an assimilated token is harder to recognise than its canonical counterpart. These results are in line with experiments on cross-modal priming of isolated words, where a single feature mismatch reduces the priming effect (Marslen-Wilson and Gaskell, 1992; Marslen-Wilson, 1993; Marslen-Wilson *et al.*, 1995). If word intelligibility effects mirror those of cross-modal priming then we would expect the assimilation of isolated word tokens to reduce the number of alveolar-final recognition responses. The loss of intelligibility for assimilated tokens in pre-velar context, in conjunction with the accompanying increase in [ŋ]-final responses reflects just such a pattern.

A preliminary examination of the intelligibility response data finds no evidence in support of the hypothesis that the asymmetry can be explained by the structure of the English Lexicon with respect to the productive -ING affix. However, we shall see, from the work presented in Chapter 10, that the relation between phonological reduction and the increased activation of lexical competitors is more important than appears at first sight.

Chapter 8

Phonological Reduction Processes II: Word-Final Stop Deletion

8.1 Introduction

The preceding chapter explored the relation between intelligibility and place assimilation of word-final nasals. It was found that running speech tends to undergo more assimilation than citation forms, and that, for words that precede velars, there is a repetition effect: nasals in second mentions are judged as more assimilated than nasals in first mentions.

In this chapter I ask whether stop deletion – like assimilation, a phonological process of reduction which characterises fast, connected speech (Shockey, 1974; Cohen and Mercer, 1975; Dalby, 1984) – might also contribute to intelligibility loss: does the deletion of word-final /d/, for example, result in fewer subjects recognising the word correctly? It is hypothesised that if the deletion of word-final stops contributes to the intelligibility loss associated with repetition, then:

Hypothesis 8.1 *Stop deletion ought to occur more frequently in running speech tokens than citation forms*

Hypothesis 8.2 *Stop deletion ought to occur more frequently for second mentions of running speech tokens than for first mentions*

Hypothesis 8.3 *an increase in stop deletion ought to correlate with a reduction in intelligibility*

Of course it is possible that the repetition effect being looked for might be manifest not in absolute deletion but in the relative terms of stop shortening. Deletion is, after all, the end-point of what is essentially a gradual (though non-linear) process of duration reduction. Consequently, the hypotheses above are adjusted to take account of possible duration effects as follows:

Hypothesis 8.4 *The duration of word-final stops will be shorter for running speech tokens than for citation forms*

Hypothesis 8.5 *The duration of word-final stops will be shorter for second than for first mentions*

Hypothesis 8.6 *The duration of word-final stops will correlate with intelligibility: the shorter the stop segment, the less intelligible the token*

These hypotheses are tested in two related studies: the first, a pilot study, investigates the incidence of both /t/- and /d/-deletion in the set of stop-final words that appeared as stimuli in the first intelligibility experiment described in Chapter 6. The second study focuses on the deletion of word-final /d/ segments, using all available data from the HCRC Map Task Corpus (Anderson *et al.*, 1991).

8.2 Pilot Study

8.2.1 Materials

The landmark names in the HCRC Map Task provide the appropriate conditioning environment for two processes of stop deletion: the deletion of /t/ in names such as *collapsed shelter* and the deletion of /d/ in landmarks like *carved stones*. An examination of the material used for the first intelligibility experiment described in Chapter 6 revealed that twelve items came from landmark names involving a possible /t/-deletion, while a further twelve items came from landmark names involving a possible /d/-deletion. These 24 items were used in a preliminary investigation into the incidence of word-final stop deletion.

8.2.2 Procedure

The intelligibility experiments had necessitated the creation of sampled speech files for each word token. The files associated with the /t/- and /d/-deletion word

forms were now segmented at a quasi-phonemic level: boundaries were drawn between all phonemic segments, and, additionally, between closure and burst phases of the word-final stop. Segmentation decisions were supported by both auditory and visual information provided by time/amplitude waveforms and wide-band spectrograms, using the ESPS XWAVES software as before. All segmentation adhered to the criteria detailed in Chapter 5 (Section 5.3). Segment durations were calculated automatically by running a simple UNIX shell script on the speech label files.

8.2.3 Results

a) /t/-deletion

	Speech form	
Mention	Citation	Token
First	5	5
Second	5	7

b) /d/-deletion

	Speech form	
Mention	Citation	Token
First	0	1
Second	0	1

Table 8.1: Incidence of /t/- and /d/-deletion for first and second mentions of citation forms and running speech tokens (N=12 in each cell)

Table 8.1 shows how frequently word-final stops were deleted in first and second mentions of both citation forms and running speech tokens. It can be seen that the incidence of /t/-deletion is high: word-final /t/ segments could not be located in almost half (5/12 = 42%) of all citation forms. An examination of the landmark names from which these words were taken revealed that the majority of /t/-deletions occurred where the stop was preceded and followed by the same segment, such as *tourist spot*, *forest stream*, *soft furnishing store*, and *nuclear test site*. Conversely, the incidence of /d/-deletion is low: only one landmark name (*diamond mine*) underwent total deletion in running speech, and in no citation forms was the /d/ fully deleted.

Given the high rate of /t/-deletion even in citation form it was decided that attention should be focused instead on the /d/-deletion dataset, since the citation forms could not be used as a reliable control for /t/-final words. The incidence

of total deletion for the /d/-final word set was low, with no sign of repetition leading to increased deletion. However, it was possible that the effects we were seeking might be found in measures of overall duration: repeated mentions might have **shorter** /d/ segments *vis-à-vis* their matched citation control than first mentions.

Analyses of Variance were therefore run on the /d/-deletion data, to investigate the effects of speech form and mention on both duration and intelligibility. In all cases mention (First/Second) and form (Citation/running speech Token) were repeated measures, with eye-contact[1] (With eye-contact/No eye-contact) as a between items grouping variable.

Separate analyses were run on closure duration only, burst duration only, and total stop duration (closure + burst). The analysis of burst duration showed no significant effects for any factor; as analyses of closure duration mirrored the main effects of overall segment duration only the latter results are reported.

A significant difference between citation forms and running speech tokens was found for word duration (Form: $F_2(1,10) = 17.59, p < .005$), segment duration (Form: $F_2(1,10) = 8.51, p = .015$), and intelligibility (Form: $F_2(1,10) = 5.08, p < .05$): citation forms were longer and more intelligible than running speech tokens, and had longer word-final /d/ segments (see Table 8.2).

However, there were no significant effects of repetition on any of the three measures, either in terms of a main effect of mention, or a form by mention interaction: word duration (Mention: $F_2(1,10) =. 4.17, p = .07$; Form x Mention: $F_2(1,10) = 1.37, n.s.$), segment duration (Mention: $F_2(1,10) = 1.70, n.s.$; Form x Mention: $F_2(1,10) < 1, n.s.$), intelligibility (Mention: $F_2(1,10) = 3.28, p = .1$; Form x Mention: $F_2(1,10) < 1, n.s.$). Neither duration – of both segment and word – nor intelligibility appear to reduce with repeated mention.

8.2.4 Conclusion

The high level of /t/-deletion in citation productions suggests that these tokens would provide inadequate controls and that extended investigations of /t/-deletion ought to be abandoned. The analysis of the twelve /d/-final items reveals a low incidence of total deletion, significant differences between citation forms and running speech tokens in duration and intelligibility, but no effects of repetition.

[1]See Anderson *et al.* (1997) for a discussion of the effects of eye-contact on speech intelligibility, where we show that first mentions in dialogues spoken With eye-contact tend to be less intelligible than introductions from No eye-contact dialogues.

Measure	Mention	Speech form Citation	Token	*loss*
WORD DURATION(ms)	First	446.8	363.1	*83.7*
	Second	430.0	313.7	*116.3*
	mean	*438.4*	*338.4*	
SEGMENT DURATION(ms)	First	78.02	53.57	*24.45*
	Second	70.14	43.15	*26.99*
	mean	*74.08*	*48.36*	
INTELLIGIBILITY	First	.601	.444	*.157*
	Second	.498	.295	*.203*
	mean	*.550*	*.370*	

Table 8.2: Mean duration (ms) of whole word and word-final /d/ segment along with intelligibility for first and second mentions of citation forms and running speech tokens (N=12)

It is conceivable, however, that the failure to find a repetition effect was a consequence of the small sample size, and that an increase in the number of data points might yield a more robust finding.

It was decided therefore to extend the analysis of word-final /d/-deletion from those items for which we had a measure of intelligibility to all suitable pairs of first and second mentions in the entire Map Task Corpus.

8.3 Main study: duration of word-final /d/

8.3.1 Materials

Since the results of the pilot analysis of the /d/-deletion word set were inconclusive, it was decided to measure the word-final stop duration for all items in the HCRC Map Task Corpus which met the criteria for inclusion detailed below. The resulting set of 54 items was consequently the maximum sample size available from this corpus of speech.

For an item to be included in the analysis, the following criteria had to be met:

1. The word-final /d/ had to occur in the **appropriate phonetic context**, that is, following a consonant in the syllabic coda and preceding a non-syllabic word onset. The original set of landmark names excluded cases

where the following word started with an oral stop because it would be difficult to locate the boundary between the unreleased /d/ and the closure period of the following stop consonant. However, to maximise the sample size this criterion was loosened and the landmark *old temple* included since it fulfilled the remaining criteria. Other than the designed set and this one addition there were no other landmark names which contained a /d/ in the appropriate phonetic context. Although it was possible that other, simply fortuitous, combinations of words within the Map Task Corpus might have resulted in a few more items which met the necessary requirements, these words would have had no citation form control and it was therefore decided not to include any such cases.

2. The landmark in question had to be introduced by either the Giver or the Follower by means of the **full referring expression** offered by the landmark label on the map; thus if *disused monastery* was introduced as simply "***a monastery***" or as "***the disused abbey-thing***" then it was excluded from further analysis.

3. The landmark then had to be referred to for a **second** time – by either speaker – again using the **full referring expression**. References to the landmark such as "*Oh, I don't have* ***that***" or "*it's okay I've got* ***the monastery*** *on mine as well*" were excluded.

4. All references had to be **fluent** and sound reasonably natural; this ruled out cases such as "*I've got* ***a gold like a mine*** *down here on the right*", for example. There was one item which sounded so unnaturally elongated that it, too, was eliminated.

5. In those cases where the Instruction Giver either introduced or repeated the landmark name, the introductory mention was deemed to be that which occurred in the **first dialogue** in which the speaker was Giver. Subsequent reference by the Giver to the same landmark in a new dialogue about the same map (but to a different Follower) was not included, on the grounds that first mentions of landmarks on second giving are less intelligible than first mentions on first giving (see Section 6.5). Since most landmarks are introduced by the Giver this restriction meant that in the majority of cases items had to be taken from conversations 1 to 4. However, self-repetition by the Follower could be taken from any dialogue, as could cases where the landmark was simply not mentioned by the Giver at all in the first

giving, making the first mention in the second giving a genuine introductory mention.

To ensure exhaustive coverage of the corpus, the original landmark reference analysis undertaken for the intelligibility experiments (see Section 6.2.1) was cross-checked with an SGML coding of the utterance, referred to here as *U coding[2]. The majority of the corpus was coded by myself and one other colleague, with the remaining few dialogues coded by two trainees whom I supervised, and whose work I subsequently checked[3].

The *U coding is essentially a more detailed version of the same reference analysis described earlier, except that it is exhaustive: every reference to any landmark is coded, with information about the landmark itself and how it was referred to. Thus a *U code will include unvarying information about whether the landmark is a shared feature, whether it provides the appropriate phonetic context for /d/-deletion to occur, and such like, in addition to information about the particular utterance, such as whether the landmark was referred to with a full or reduced referring expression, whether the reference was the first or a repeated mention, whether the form of referring expression was definite, indefinite, pronominal or deictic, and whether it was disfluent. Example (8.1) contains the *U coding associated with the first part of the dialogue excerpt in Section 5.2.6.1. Details about the codes themselves can be found in Appendix C.

(8.1)
```
TA:  Start at the extinct volcano,
*U in extinctjjvolcano tdel same intro men def l
     and go down round the tribal settlement.  And then
*U in dir tribaljjsettle SW dif f01 intro men def l

TB:  Whereabouts is the tribal settlement?
*U qu loc tribaljjsettle SW dif f01 rep def l

TA:  It's at the bottom.
*U loc tribaljjsettle SW dif f01 rep pro
It's to the left of the e extinct volcano.
*U loc tribaljjsettle SW dif f01 rep pro
   dir extinctjjvolcano tdel same rep def ld
```

[2]This coding system was originally devised in Glasgow for use with Map Task dialogues by school children, and, with some revision, was then applied to the HCRC Map Task Corpus

[3]The *U coding has since undergone further revision by M. Aylett.

```
TB:  Right.  How far?
*U resp + qu dis

TA:  Ehm, at the opposite side.
*U qresp loc

TB:  To the opposite side.  Is it underneath the rope bridge or
     to the lef
*U qloc tribaljjsettle SW dif f01 rep pro
   dir ropejjbridge same intro men def l

TA:  It's underneath the rope bridge.
*U qresp loc tribaljjsettle SW dif f01 rep pro
   dir ropejjbridge same rep def l
```

The *U coding did not exist at the time of running the intelligibility experiments, but its subsequent availability had two obvious advantages: first it meant that the original paper-based analysis could be checked against the on-line coded version and any discrepancies brought to light. This checking process highlighted one or two coding errors, as well as one omission in the paper version where a reference had been overlooked (this error was irrelevant to the intelligibility work done earlier). Secondly, the coding made it possible to locate and extract automatically all literal introductions and repeated mentions of potential /d/-deletion landmark names. After the output had been screened for Giver mentions in later dialogues (see above) a total of 54 items remained to be measured and analysed, of which 17 were unique word forms, that is, different lexical items.

8.3.2 Procedure

The utterance containing each referential expression was extracted from the digitised speech available via the CD-ROM. The same segmentation procedure was used as that for the pilot study. In the case of *old temple* the duration of the word-final /d/ was combined with the closure period of the following /t/ since it was not always possible to locate a boundary between the two. In this way measurements would at least be consistent, and comparable across conditions. To test whether the inclusion of the /t/ duration made a significant contribution to the results reported below, separate analyses were run on the data with *old*

temple excluded. The analyses conformed on all the critical results, suggesting that the exclusion of *old temple* was unnecessary.

8.3.3 Results

With the set of items now extended to cover all first and second mentions from the HCRC Map Task Corpus whether or not they were used in one of the intelligibility experiments, it is not possible to test for an effect of repetition on intelligibility for the dataset as a whole. Given that the current goal is to ascertain whether word-final stop deletion is a likely source of intelligibility loss from first to second mention, what is required is some alternative measure of reduction to confirm that there is a repetition effect to be accounted for in this set of 54 word items. Such a measure is provided by a word's k-score.

It has been argued (Campbell and Isard, 1991) that absolute word duration fails to take account of the inherent variation in the tendencies of individual phonemic segments to expand and contract. While some segments are relatively elastic, other segments have a less flexible duration which does not fluctuate much between productions. As Campbell and Isard observe:

> *"A segment with a high variance (such as a tense vowel) that shows considerable difference in duration in terms of absolute millisecond measurements may be in the same relative state of expansion or compression as one with a much smaller variance (a stop, for example) that appears to change less in absolute terms." (1991:40)*

As a means of accounting for some of the variation in duration associated in particular with the effects of rhythmic stress, Campbell and Isard offer a measure of segment **elasticity**, where elasticity is calculated in relation to a token's z-score: the measure of a token's distance, in standard deviations, from the mean duration for that segment type. The k-score of a given syllable is then computed by calculating the **average** z-score of the phonemes that make up the syllable. The strong form of the elasticity hypothesis states that:

> *"all segments in a given syllable fall at the same place in their respective distributions. That is, for any given syllable, there should be a number k of standard deviations such that the length of every segment in the syllable is equal to* $\mu_{seg} + k\sigma_{seg}$*, where* μ_{seg} *and* σ_{seg} *are the mean and standard deviation respectively of durations of the particular segment type." (Campbell and Isard, 1991:40)*

The k-score of a syllable can thus be defined as:

$$k_{syll} = \frac{dur_{syll} - \sum_{i=1}^{n} \mu_i}{\sum_{i=1}^{n} \sigma_i}$$

where:

- dur_{syll} is the observed total duration of the syllable;
- μ_i is the mean of the i^{th} phoneme in the syllable;
- σ_i is the standard deviation of the i^{th} phoneme in the syllable;
- n is the number of phonemes in the syllable.
  (Molloy, 1997)

Generalising the principle of k-score duration from syllable to word length units (Molloy, 1997, personal communication), k-score duration was calculated for all 54 words[4] in the /d/-final word set. An Analysis of Variance on k-score duration, grouping by the between items factor eye-contact (With/No), with speech form (Citation/running speech Token) and mention (First/Second) as repeated measures, revealed significant main effects of form ($F_2(1, 52) = 120.03, p < .0001$) and mention ($F_2(1, 52) = 4.42, p < .05$), and a significant form by mention interaction ($F_2(1, 52) = 5.45, p < .05$) (see Table 8.3). A *post hoc* Scheffé test on the interaction shows the difference in k-score between first and second mentions of running speech tokens to be significant ($p < .05$) while k-score for citation forms does not differ by mention. Repeated mentions, then, are more compressed than introductory mentions of the same word.

	Speech form		
Mention	Citation	Token	***loss***
First	0.980	0.488	*.492*
Second	0.977	0.358	*.619*
mean	*.979*	*.423*	

Table 8.3: Mean k-score duration ($s.d.$s) for all words in the /d/-final dataset (N=54), looking at first and second mentions of citation forms and running speech tokens

When the phonetic identity of the segments that compose the word are taken into account, in conjunction with the tendency each segment has to expand and

[4]Since many of the /d/-deletion items were monosyllabic the difference between k_{syll} and k_{word} should not be great

contract, a significant effect of repetition is found: second mentions contract more than first mentions. This result suggests that had we access to intelligibility values for all 54 items we would probably find a similar effect of repetition on intelligibility. For this reason it is appropriate to explore the incidence of word-final /d/-deletion, to see whether a similar effect of repetition can be found.

A tally of the incidence of full stop deletion – that is, the total absence of any stop-like characteristics in the speech waveform – shows that while only one /d/ segment was fully deleted from citation form readings, there were eleven deletions for first mentions of landmarks taken from the non-scripted dialogues, and twelve deletions for second mentions (see Table 8.4). As in the pilot study, there is evidence to suggest that the production of running speech tokens involves the more frequent application of phonological reduction processes than does the production of citation forms, thus confirming Hypothesis 8.1; however, the incidence of outright deletion again fails to demonstrate an effect of repetition within the running speech tokens (Hypothesis 8.2).

	Speech form	
Mention	Citation	Token
First	1	11
Second	0	12

Table 8.4: Incidence of /d/-deletion for first and second mentions of citation forms and running speech tokens (N=54)

An examination of the identities of the landmarks involving full /d/-deletion revealed that the landmark name most likely to lose its word-final /d/ was *diamond mine*: 14 of the 24 items involving a full stop deletion (=58%) were tokens of *diamond mine*. Although this was the most frequently occurring landmark name in the /d/-deletion dataset (because of its status as a master feature), it only contributes about 20% (10/54) of the items analysed, a far smaller proportion than the 58% of deletions to which it contributes.

Since deletion is the end point of a gradual process of duration reduction, a 2x2 Analysis of Variance was performed on the **raw duration** (ms) of the word-final stop segments, with speech form (Citation/running speech Token) and mention (First/Second) as repeated measures, and eye-contact (With eye-contact/No eye-contact) as a between items grouping variable (see Anderson *et al.*, 1997).

As can be seen from the values in Table 8.5 although stops are longer in citation form than running speech (Form: $F_2(1, 52) = 35.16, p < .0001$), there is no main

a) Segment duration (ms)

	Speech form		
Mention	Citation	Token	***loss***
First	79.30	43.76	*35.54*
Second	81.77	42.18	*39.59*
mean	*80.54*	*42.97*	

b) Word duration (ms)

	Speech form		
Mention	Citation	Token	***loss***
First	409.1	309.9	*99.2*
Second	407.0	290.5	*116.5*
mean	*408.1*	*300.2*	

Table 8.5: Mean duration (ms) of word-final /d/ segment and whole word for first and second mentions of citation forms and running speech tokens (N=54)

effect of mention (Mention: $F_2(1, 52) < 1$) nor any interaction (Mention x Form: $F_2(1, 52) < 1$): word-final stops in second mentions are no shorter than in first mentions, either for citation forms or running speech tokens. The results hold true whether the stop is measured as closure duration only, burst duration only, or the sum of the two (i.e. full stop duration). Thus there is evidence in support of a citation/token difference (Hypothesis 8.4) but not for an effect of repetition (Hypothesis 8.5).

Table 8.5 also contains the mean duration (ms) for the **words** from which the stop measurements were taken. A 2x2 Analysis of Variance, again with form (Citation/Token) and mention (First/Second) as repeated measures, and eye-contact (With/No) as a grouping factor, but this time with the duration of the whole word as the dependent variable, revealed a similar form effect: citation forms are longer than running speech tokens (Form: $F_2(1, 52) = 120.71, p < .0001$). However, in the analysis of word duration, there is now evidence of a trend towards a repetition effect, with both the main effect of mention and the mention by form interaction approaching significance (Mention: $F_2(1, 52) = 3.81, p = .056$; Form x Mention: $F_2(1, 52) = 2.83, p = .0988$). Recall that when the tendency of particular phonemes to expand and contract is taken into consideration (k-score), the effect of repetition reaches statistical significance.

An ANOVA was also run on the **reduction from citation** of word duration, that is, using the change in duration from citation form to running speech token as the dependent variable, with mention as a repeated measure, and eye-contact

as a between items factor. Looking at the means in Table 8.6, it is apparent that

	Eye-Contact		
Mention	No Eye	With Eye	***mean***
First	76.6	123.6	*99.2*
Second	110.8	122.7	*116.5*
diff	*34.2*	−0.9	

Table 8.6: Mean duration difference from citation (ms) for whole words in /d/-deletion dataset, comparing first and second mentions of landmark names taken from dialogues where speakers either could (With Eye, N=26) or could not (No Eye, N=28) see their partner's face

while there is an effect of repeated mention for the data from the No eye-contact condition, words from the With eye-contact dialogues show no repetition effect. This difference approaches statistical significance (Mention x Eye: $F_2(1, 52) = 3.14, p = .082$). An analysis of the No eye-contact data alone, taking reduction from citation as the dependent variable and mention as a repeated measure, reveals significantly greater reduction for second mentions than first mentions (Mention: $F_2(1, 26) = 6.02, p < .05$).

On the basis of this eye-contact effect the stop duration measurements were re-analysed looking at the No eye-contact cases separately. When the With eye-contact data are excluded, the new 'reduction from citation' analysis only raises the original F value for mention to a fraction over 1 (Mention: $F_2(1, 26) = 1.38, p = .25$). While the means are in the predicted direction the standard deviation is large: there is clearly too much variability to get a significant result. So although the No eye-contact data are a little more in line with our predictions than the full set of data the basic finding remains that repetition has no significant effect on word-final stop duration.

8.3.3.1 Word-final /d/ and intelligibility

For those cases where we do have values for intelligibility, it should be possible to explore the relation between intelligibility and duration. The pilot study investigated the 12 /d/-deletion items that appeared in Experiment One. The remaining three intelligibility experiments yielded only 6 more sets of tokens (where a set consists of the first and second mention with corresponding citation forms), providing a total of 18 items. The total group included 13 unique landmark names and 5 cases of duplication, arising from the presence of the same landmark (such as *diamond mine*) on more than one map pair. Recall

that a relation was predicted between intelligibility loss and the duration of word-final /d/: loss of intelligibility should correlate with deleted, or shorter, /d/ segments. Although there are significant correlations between intelligibility and word duration ($N = 88, r = .412, p < .001$) and intelligibility and k-score ($N = 88, r = .416, p < .001$) – more intelligible words are longer and more stretched than less intelligible words – there is no direct relation between intelligibility and word-final stop duration ($N = 88, r = -.124, p = .248, n.s.$).

A subsequent set of correlations looked at the reduction in intelligibility from citation to running speech token in relation to the duration of the token's word-final stop. While intelligibility loss failed to correlate with token stop duration for first mentions ($N = 18, r = .257, p = .304$), there was a significant effect for second mentions ($N = 18, r = .462, p = .05$). However, the result is in the reverse direction to the prediction above, with longer stops correlating with greater intelligibility loss.

The reason for this initially counter-intuitive finding becomes apparent when we look at the individual cases: three items have completely deleted stops but are nevertheless highly intelligible in both running speech and citation form (with 'reduction from citation' therefore small or zero); a further three items – those with the shortest stops – are completely unintelligible in both citation and running speech form (again resulting in no reduction). Thus for the cases with no or very short stops there is no reduction from citation form for intelligibility because either t2 and c2 are both high, or t2 and c2 are both low (where t2 represents the intelligibility of the second token from running speech, while c2 represents the intelligibility of the corresponding citation form).

This observation serves as a reminder that word intelligibility is not simply a matter of sequential segment perception, but depends on the size and nature of a target word's competitor set. The auditory illusion of phonemic restoration (Samuel, 1981, 1987) illustrates how a word can still be recognised when part of it is replaced by noise. The redundancy in the signal means that not all the information available is required. There may be so few competitors to the word *diamond*, for example, that even when the final /d/ is entirely deleted, subjects have no difficulty in recognising the word. Conversely, the word *owed* may be difficult to recognise even with a fully articulated final /d/ because it is short and because there are many similar sounding words with which it has to compete, such as *owned, old* and *odour.*

Recall that at the start of this chapter the following two hypotheses were osten-

sibly equated:

- an increase in /d/-deletion ought to correlate with a reduction in intelligibility.
- duration of word-final /d/ will correlate with intelligibility: the shorter the stop segment, the less intelligible the token.

However, the lexical consequences of /d/-deletion and duration reduction – or /d/-shortening – are quite different: the actual deletion of a segment can potentially change the lexical competitor set in a way that shortening does not.

In Chapter 10 I pursue this idea further, by looking at the error responses offered by the intelligibility subjects as alternative candidates to the target word. I consider what effects the deletion of the word-final segment may have on the set of competitors that gets activated and ask if the deletion of word-final /d/ segments depends on whether it belongs to the word stem or the past tense -ED affix.

8.4 Discussion

The analysis using k-scores reveals that there *is* a discourse effect of repetition: productions that introduce entities into the discourse are less compressed than productions which refer back to an entity mentioned previously. However, the repeated mention effect is not replicated in the rate of /d/-deletion, nor in measures of word-final /d/ duration: the incidence of deletion in running speech tokens is the same for both first and second mentions, while /d/ durations in second mentions are no shorter *vis-à-vis* citation form than /d/ durations in first mentions. Rather, the reduction which takes place at the end of the word fails to reflect any difference between introductory and repeated mentions. In other words, although there was evidence in support of a speech form effect (Hypotheses 8.1 and 8.4) there was no support for any effect of repetition on either deletion (Hypothesis 8.2) or duration (Hypothesis 8.5), nor was a relation found between the reduction of word-final /d/ segments and intelligibility (Hypotheses 8.3 and 8.6).

There are at least two reasons why this might be. Consider, first, the findings of Campbell and Isard (1991) in relation to segment elasticity: while segments with high variance include tense vowels, the stops being examined here fall into the category of 'low variance' segments: if the duration of /d/ has a low mean,

and a relatively small standard deviation, then it may not be possible to find significant differences in duration across what are essentially quite small numbers of items. Perhaps if, instead, we examined the duration of the stressed vowel – which, according to Campbell and Isard, ought to demonstrate greater variance – we might find a significant effect of repetition.

Given that the /d/-deletion stimuli had been segmented through the whole word at a phonemic level, it was possible to test this hypothesis directly by subjecting the duration of the stressed vowels to an Analysis of Variance similar to those run on word-final /d/ duration. Items were grouped by eye-contact (With/No), with speech form (citation/Token) and mention (First/Second) as repeated measures. Mean duration values can be found in Table 8.7. Stressed vowels are longer in

	Speech form		
Mention	Citation	Token	***loss***
First	167.5	130.9	*36.6*
Second	167.7	117.7	*50.0*
mean	*167.6*	*124.3*	

Table 8.7: Mean duration (ms) for all stressed vowels in the /d/-final dataset (N=54), looking at first and second mentions of citation forms and running speech tokens

citation form than running speech (Form: $F_2(1,52) = 75.52, p < .0001$) and shorter for second than for first mentions (Mention: $F_2(1,52) = 5.61, p < .05$). The shortening from first to second mention is revealed to hold only for running speech tokens and not for the citation form controls (Form x Mention: $F_2(1,52) = 8.15, p < .01$; Scheffé: $c1 = c2, c1 > t1, c2 > t2, t1 > t2$ (all at $p < .01$)). Thus, while we fail to find an effect of repetition on the duration of /d/ segments – segments which demonstrate only a small variance in duration – we do find an effect on the duration of stressed vowels – segments characterised by much greater variance – with vowels shortening significantly in repeated tokens.

Recall that whilst we have observed significant effects of repetition on intelligibility (and, now, on stressed vowel duration), we have failed to replicate repetition effects both for word-final /d/-deletion, and for the duration of word-final /d/ segments. The failure to find effects of word-final place assimilation of /n/ preceding labial contexts in the previous chapter lends further weight to the suggestion that these phonological reduction processes have little effect on intelligibility, **because they occur late in the word**. The reason why stressed vowels show the effect of repetition while word-final stops do not may be related to the **location** of the

segment within the word, rather than the segment's natural compressibility *per se*.

Intelligibility scores reflect subjects' ability to recognise words, that is, to match the acoustic input to an internally stored lexical representation. Phonological reduction processes that operate at the ends of words may well be acting on segments that contribute little to recognition effects: if the deletion of word-final /d/ occurs after a word is uniquely identifiable then the loss of the /d/ segment may make little difference to the subjects' ability to recognise the word: the word may already have been recognised. Changes to the production of the stressed vowel, on the other hand, have a greater likelihood of affecting recognition responses, since identifying the vowel is crucial to most traditional theories of lexical access (Forster, 1976; Marslen-Wilson and Tyler, 1980; Gaskell and Marslen-Wilson, 1996). Perhaps, then, we should be looking **earlier** in the word – before a word's Uniqueness Point, for example – for a suitable candidate for the source of intelligibility loss associated with repetition? In the next chapter, therefore, we direct our attention to a phonological reduction process that occurs early in the word: pre-stress schwa syncope. We ask whether the incidence of schwa-syncope in Weak initial syllables – such as [sə] in *saloon* – increases with repetition. We will also consider the effects of repeated mention on the duration of stressed vowels for the full set of materials used to investigate phonological reduction.

Chapter 9

Phonological Reduction Processes III: Vowel duration

9.1 Introduction

The previous two chapters considered the effects on intelligibility of two word boundary reduction processes: place assimilation and stop deletion. The contribution made by these two processes to the intelligibility loss observed for reference repetition was found to be small: while assimilation to velar place *was* associated with a reduction in intelligibility, there was no effect of repetition on assimilation to labial place, nor on either the deletion or shortening of word-final stop consonants. I suggested that a possible explanation for these results lay in the *location* of the reduction process at the word boundary: if reduction occurs beyond the point at which a lexical item becomes unique, then the effects of such reduction on subjects' ability to recognise the word may be small.

In this chapter, therefore, I focus on a phonological reduction process that occurs much earlier in the production of a word: schwa syncope in polysyllabic words with metrically Weak initial syllables. The initial set of hypotheses are similar to those for /d/-deletion in the previous chapter; if the deletion of schwa in metrically Weak onsets contributes to the intelligibility loss associated with repetition, then:

Hypothesis 9.1 *Schwa-syncope ought to occur more frequently in running speech tokens than citation forms*

Hypothesis 9.2 *Schwa-syncope ought to occur more frequently for second mentions of running speech tokens than for first mentions*

Hypothesis 9.3 *An increase in schwa-syncope ought to correlate with a reduction in intelligibility*

The H & H theory predicts greater hypo-articulation in repeated mentions than introductory mentions, in other words, an increase in articulatory economy with repetition. This means that speakers could shorten schwa articulation without fully deleting the vowel. Therefore a further set of hypotheses were formulated with respect to the **duration** of pre-stress schwa:

Hypothesis 9.4 *The duration of schwa in the initial syllable of WS polysyllables will be shorter for running speech tokens than for citation forms*

Hypothesis 9.5 *The duration of schwa will be shorter for second than for first mentions*

Hypothesis 9.6 *Schwa duration will correlate with intelligibility: the shorter the weak vowel, the less intelligible the token*

An appropriate test of these hypotheses requires an analysis of schwa in metrically Weak initial syllables of polysyllabic words that are mentioned twice or more in spontaneous discourse: the HCRC Map Task Corpus (Anderson *et al.*, 1991) provides just such a set of materials.

9.2 Duration of schwa in WS polysyllables

9.2.1 Materials

Recall that the landmark names used to label map features in the HCRC Map Task Corpus (Anderson *et al.*, 1991) were selected on the basis of their phonological characteristics (see Chapter 5 Section 5.2.2.1). In addition to feature names involving word boundary reduction processes, every map contained an example of each of two **polysyllabic** word categories, based on the metrical structure of the word-initial syllable. Polysyllables were:

- word-initial Strong-Weak (SW) words e.g. *buffalo, monastery*
- initial Weak-Strong (WS) words e.g. *baboons, machete*

To test whether repetition leads to an increase in schwa syncope and/or a reduction in duration and intelligibility, **pairs of first and second mentions** are required, where each is a **full literal mention** of the landmark name as labelled on the map. The relation between repetition and intelligibility is tested by running appropriate Analyses of Variance on the set of polysyllabic landmark names which appeared in the intelligibility studies described in Chapter 6. The effect of repetition on schwa syncope is tested by labelling the component phonemic segments in each WS polysyllable, looking for evidence of deletion, and running Analyses of Variance on the duration of those weak vowels which are still articulated. Correlational statistics are used to test for a relation between schwa duration and intelligibility.

The full set of maps (16 pairs) offers a total of 81 landmark names which contain words of two or more syllables with either SW or WS word onsets. Of these 81 unique items, 59 were found to have been used in one or more of the intelligibility experiments described in Chapter 6. Because the same maps were used in several dialogues with different speakers some of these landmark names appeared more than once within an experiment[1]. Likewise, a set of tokens (first and second mentions with matched citation forms) was occasionally reused in a subsequent experiment where the token set met the appropriate conditions of sharedness, feedback, speaker ID *etc.* Consequently, the 59 unique word forms appeared in a total of 90 polysyllabic token sets which had been presented to subjects and for which there were intelligibility scores. Of these, 27 of the token sets had Weak onsets, while the remaining 63 had Strong onsets.

9.2.2 Procedure

The same procedure was used as that for the pilot study in Chapter 8 (see Section 8.2.2).

9.2.3 Results

9.2.3.1 Replicating the repetition effect

Given that polysyllabic words are easier to recognise in a lexical identity task than monosyllables (Craig and Kim, 1990), it was important to ascertain whether there

[1]Recall that in all but one of the experiments two sets of materials were used so that word forms could be matched across experimental conditions (such as same versus different repeater); duplication also encouraged maximum inclusion of available material.

was a ceiling effect for the intelligibility of the polysyllable dataset: were all polysyllabic tokens so easy to recognise that there was no advantage for citation forms over running speech tokens, or for first over second mentions? A 2x2 Analysis of Variance was therefore performed on the intelligibility scores for the full set of 90 items, with form (Citation/running speech Token) and mention (First/Second) as independent variables within items, and metrical structure (SW/WS) as a grouping variable between items. Table 9.1 shows the mean intelligibility for first and second mentions compared with their matched citation forms, while Table 9.2 shows the mean intelligibility differences for SW and WS polysyllables. Table 9.3 depicts the intelligibility loss from citation for first and second mentions of both SW and WS polysyllables.

	Speech form		
Mention	Citation	Token	***loss***
First	.832	.721	*.111*
Second	.804	.559	*.245*
mean	*.818*	*.640*	

Table 9.1: Mean intelligibility for polysyllables comparing first and second mentions from spontaneous dialogue with matched citations (N=90)

Metrical	**Speech form**		
structure	Citation	Token	*N*
SW	.839	.703	*63*
WS	.769	.492	*27*
mean	*.804*	*.598*	

Table 9.2: Mean intelligibility for polysyllables with Strong- or Weak-initial syllables, comparing tokens produced in unscripted dialogue with matched citations

	Metrical structure		
Mention	SW	WS	*Mean*
First	.079	.187	*.111*
Second	.194	.367	*.246*
N	*63*	*27*	*90*

Table 9.3: Mean intelligibility loss from citation form, for first and second mentions of polysyllables with SW or WS onsets

Citation forms are significantly more intelligible than running speech tokens (Form: $F_2(1,88) = 40.57, p < .0001$), with intelligibility also reducing significantly from first to second mention (Mention: $F_2(1,88) = 12.59, p < .001$). A

post hoc Scheffé test on the significant form by mention interaction ($F_2(1,88) = 8.04, p < .01$) revealed no difference between citation forms, while second mentions were significantly less intelligible ($p < .01$) than first mentions; running speech tokens were less intelligible than their citation match (first mentions: $p < .05$; second mentions: $p < .01$). In other words, the repetition effect is replicated for the subset of polysyllabic words, despite the fact that many of these word forms were more than two syllables long[2].

The effect of metrical structure was also found to be significant: words with Weak-initial syllables are less intelligible than words with Strong syllable onsets ($F_2(1,88) = 8.83, p < .005$). There was also a significant interaction between metrical structure and form ($F_2(1,88) = 4.70, p < .05$), with a Scheffé test indicating that WS tokens from unscripted dialogue are less intelligible than SW tokens ($p < .05$), but that citation forms do not differ from each other, although they are more intelligible than their matched running speech forms (SW: $p < .05$; WS: $p < .01$).

The question remains whether a repetition effect holds for the WS polysyllabic data alone. A further 2x2 ANOVA was run, this time on the Weak-initial dataset only (N=27), again with form (Citation/Token) and mention (First/Second) as repeated mentions, but with eye-contact (With/No) as the grouping factor. As

	Speech form		
Mention	Citation	Token	***loss***
First	.758	.571	*.187*
Second	.779	.413	*.366*
mean	*.769*	*.492*	

Table 9.4: Mean intelligibility for Weak-initial polysyllables comparing first and second mentions from spontaneous dialogue with matched citations (N=27)

can be seen from the figures in Table 9.4 citation forms are more intelligible than running speech tokens (Form: $F_2(1,25) = 31.91, p < .0001$), while second mentions suffer greater intelligibility loss *vis-à-vis* matched citations than do first mentions (Form x Mention: $F_2(1,25) = 4.22, p = .05$). However, a *post hoc* Scheffé test reveals that although running speech tokens are less intelligible than their matched citations (first mention: $p < .05$; second mention: $p < .01$) the difference in intelligibility between citations and between first and second men-

[2] A subsequent analysis, grouping by number of syllables, did reveal a predictable significant effect of syllable length: two syllable words were less intelligible than three syllable words (mean intelligibility: bisyllables=.688 (N=60); trisyllables=.820 (N=28); $F_2(1,86) = 7.77, p < .01$).

tions of running speech tokens does not quite reach significance (critical value for $p < 0.05 = 4.87$, citations = 0.58; tokens = 4.27). There are no effects of eye-contact (Eye: $F_2 < 1$).

9.2.3.2 Schwa duration

Dalby found that in his corpus of 'television English' roughly 9% of unstressed vowel environments occurred with no audible schwa (1984:16), while the incidence of schwa deletion rose to 43% in a corpus of 'fast' read speech (1984:41). In the data examined here, evidence of pre-stress syncope is minimal, with just one token exhibiting total deletion (see Table 9.5). With respect to the first three

	Speech form	
Mention	Citation	Token
First	0	0
Second	0	1

Table 9.5: Incidence of schwa-syncope in WS polysyllables, comparing first and second mentions of citation forms and running speech tokens (N=27)

hypotheses (9.1-9.3), the only observations that can be made are that this one case of syncope did occur in running speech rather than citation form, and in a second rather than first mention. The token that underwent syncope was also more difficult to recognise than either its citation, or the first mention by the same speaker (intelligibility for token1=0.9, cit1=1, token2=0.6, cit2=1). Given the failure to find more than one case of genuine deletion, it was necessary to turn to the set of hypotheses relating to the **duration** of the articulated schwa.

Measurements of schwa duration were subjected to a 2x2 Analysis of Variance with form (Citation/running speech Token) and mention (First/Second) as repeated measures, and eye-contact (With/No) as a between items grouping factor. Mean duration values can be found in Table 9.6. Although the schwa in citation forms is significantly longer than in running speech tokens (Form: $F_2(1, 25) = 17.5, p < .0005$), there is no effect of repetition (Mention: $F_2 < 1$; Mention x Form: $F_2 < 1$). Indeed, the means for segment duration in Table 9.6 fail even to fall in the predicted direction. There is no main effect of eye-contact (Eye: $F_2 < 1$), but a significant form by eye-contact interaction ($F_2(1, 25) = 4.32, p < .05$) is shown, by a *post hoc* Scheffé test, to reflect a failure for tokens from No eye-contact dialogues to demonstrate a duration difference between citation forms and running speech; the form difference for tokens spoken

with eye-contact is significant ($p < .01$). Neither citations nor tokens were found to differ across eye-contact conditions.

a) Segment duration (ms)

	Speech form		
Mention	Citation	Token	***loss***
First	66.13	51.42	*14.71*
Second	64.79	52.80	*11.99*
mean	*65.46*	*52.11*	

b) Word duration (ms)

	Speech form		
Mention	Citation	Token	***loss***
First	605.7	457.4	*148.3*
Second	603.9	405.8	*198.1*
mean	*604.8*	*431.6*	

Table 9.6: Mean duration (ms) of whole word and of schwa segment in Weak-initial polysyllables, comparing first and second mentions from spontaneous dialogue with matched citations (N=27)

A 2x2 Analysis of Variance was also run on the mean **word** duration for the WS polysyllables from which the schwa measurements were taken, again with form (Citation/Token) and mention (First/Second) as repeated measures, and eye-contact (With/No) as a grouping factor (see Table 9.6 for means). There are significant effects on overall word duration for both form (Form: $F_2(1,25) = 69.98, p < .0001$) and mention (Mention: $F_2(1,25) = 9.60, p < .005$), as well as a significant interaction (Form x Mention: $F_2(1,25) = 8.72, p < .01$): citation forms are longer than running speech tokens, and first mentions are longer overall than second mentions. A *post hoc* Scheffé test on the interaction reveals that though citation forms are longer than running speech tokens, the significant difference between first and second mentions holds for running speech tokens only ($p < .01$); citation forms do not differ across mention[3].

9.2.3.3 Schwa duration and intelligibility

The effect of duration on intelligibility was tested by a series of correlations. A simple correlation between schwa duration and intelligibility for all tokens

[3]Although there is a significant interaction between form and eye-contact (Form x Eye: $F_2(1,25) = 6.93, p < .05$), the effect of eye-contact is shown by Scheffé not to be significant, with no difference between tokens in the two eye-contact conditions, or between citation forms.

showed that polysyllables with shorter weak vowels were significantly less intelligible ($N = 108, r = .313, p < .001$). However, such an analysis does not control for the effects of speech form: the significant correlation might simply arise from the fact that citation forms are longer and more intelligible than running speech tokens. To explore the relation between schwa duration, intelligibility and repetition, therefore, we need to relate the intelligibility loss (from citation to token) associated with first and second mentions with corresponding changes in duration. The relation between intelligibility loss ($intell_{cit} - intell_{tok}$) and duration difference ($dur_{cit} - dur_{tok}$) for the schwa segment approaches significance for first mentions ($N = 27, r = .364, p = .062$) and is significant at $p < .05$ for second mentions ($N = 27, r = .428, p = .026$). The equivalent analyses for word duration are not significant (first mentions: $N = 27, r = .166, p = .41$; second mentions: $N = 27, r = .281, p = .16$), despite the significant relation between overall shortening ($durW_{cit} - durW_{tok}$)[4] and the shortening of pre-stress schwa (first mentions: $N = 27, r = .558, p < .005$; second mentions: $N = 27, r = .636, p < .001$). It would appear, then, that there is a weak relation between intelligibility loss and pre-stress schwa duration, but no relation between loss of intelligibility and overall word duration. The relation between intelligibility loss and weak vowel duration change is stronger for second mentions than for first mentions.

9.2.4 Discussion

The incidence of complete schwa syncope in the set of WS polysyllables that were used in the intelligibility experiments was low: in only one token was there evidence of total deletion. Therefore Hypotheses 9.1-9.3 could not be addressed. Analyses of segment duration confirmed Hypothesis 9.4 that schwa would be shorter in running speech tokens than in citation forms. There was no support for Hypothesis 9.5: repeated mention of WS polysyllables had no effect on the duration of the weak vowel, although analyses of overall word duration did reveal a significant effect of repetition. A significant relation was found, however, between the duration of schwa and word intelligibility; intelligibility loss was also found to correlate with reduction in weak vowel duration. These results appear to confirm Hypothesis 9.6.

The lack of a significant repetition effect on schwa duration might be accounted for by a 'floor' effect: it may be that the weak vowel is already so short in the introductory mention that it cannot get significantly shorter when repeated. How-

[4] where $durW$ stands for overall word duration

ever, such an explanation fails to account for why speakers should resist deleting the schwa altogether. Dalby (1984)'s findings suggest that schwa syncope occurs most frequently in words where the remaining consonants can resyllabify into an acceptable cluster of English; thus speakers might delete the schwa in *saloon* but not in *ravine* or *machete*. Of the 27 token sets used in the analysis above, only 6 offer acceptable onsets when schwa is deleted (*saloon*×3, *allotments*×2, *apache*). This may go some way to explaining speakers' tendency to articulate some sort of weak vowel.

It is worth noting that other researchers have observed that duration measures are less likely to show significant effects of information value (Given/New) than, for example, spectral analyses (van Bergem and Koopmans-van Beinum, 1989). Measuring target undershoot for weak vowels like schwa is not without its problems, however: there is evidence to suggest that schwa may in fact be target-less (Bates, 1995), in which case it is not clear what kind of spectral control ought to be used. Of course the effects of variability in duration can be reduced by increasing the sample set; however, increasing the number of tokens measured will not produce a significant repetition effect if durations are at minimal values even for first mentions.

One of the motivations for exploring pre-stress syncope was to locate an effect of repetition on a phonological reduction process which occurs early in a word's production, that is, before the lexical item becomes uniquely identifiable. No repetition effect was found for the duration of pre-stress schwa in WS polysyllables, but it may be that the opportunity for schwa to expand or contract in these Weak syllables is limited. In the previous chapter an effect of repetition *was* observed for the duration of *stressed* vowels, which are known to exhibit a high level of elasticity (Campbell and Isard, 1991). In the following section, therefore, the study of stressed vowel duration is extended to cover the full set of materials used in the three studies of phonological reduction: the nasal data, the /d/-deletion data and the polysyllable data.

9.3 Duration of stressed vowels

9.3.1 Materials and Procedure

All materials from the nasal, /d/-deletion and WS polysyllable studies were pooled to provide a total of 114 (33+54+27) sets of tokens, where each set con-

sisted of the first and second mentions of the landmark name with corresponding citation forms. Since all tokens had already been segmented at a phonemic level, the duration of the stressed vowel could be calculated by running a UNIX shell script on the individual speech label files and extracting the relevant value. Duration values were entered into a data matrix along with intelligibility scores where available and subjected to a series of ANOVAs.

9.3.2 Results

An Analysis of variance was conducted on raw vowel duration with speech form (Citation/running speech Token) and mention (First/Second) as repeated measures, and both phonological reduction category (Nasal/WSpoly/d-del) and eye-contact (With/No) as grouping factors between items. The mean duration values can be found in Table 9.7.

	Speech form		
Mention	Citation	Token	***loss***
First	147.8	116.8	*31.0*
Second	148.7	108.2	*40.5*
mean	*148.25*	*112.5*	

Table 9.7: Mean duration (ms) for all stressed vowels from full set of materials (N=114), looking at first and second mentions of citation forms and running speech tokens

Significant main effects are found for form, eye-contact and phonological reduction category. Vowels in citation forms are longer than vowels in words from running speech (Form: $F_2(1, 108) = 119.29, p < .0001$); vowels in words from dialogues with no eye-contact are longer than vowels from dialogues where speakers are able to see each other's faces (Eye: $F_2(1, 108) = 6.58, p = .01$)[5]; vowels are longer in the words from the /d/-deletion materials than those in either the nasal or WS polysyllable materials (Phon: $F_2(2, 108) = 8.49, p < .0005$; Scheffé: ddel>nasal ($p < .01$), ddel>WSpoly ($p < .05$), nasal=WSpoly).

[5] Although the significant effect of eye-contact appears at first sight to mirror the intelligibility loss reported in Anderson *et al.* (1997) for introductory mentions from eye-contact dialogues, there is no significant interaction between eye-contact and form for the vowel duration reported here: vowels are longer in the face-screened condition for both running speech tokens and for citation forms. Given that the material in Anderson *et al.* was carefully matched across eye-contact condition while the material here was not, it is likely that the duration difference observed above arises from differences in vowel type and/or word length between the two sets of materials rather than face visibility *per se*.

Analysis of Variance also revealed a significant interaction between form and mention, with similar durations for both citation conditions but with longer vowels in spontaneous introductions than in repeated mentions (Form x Mention: $F_2(1, 108) = 7.84, p < .01$; Scheffé: c1=c2; c1>t1; c2>t2; t1>t2 (all at $p < .01$)).

A similar ANOVA was run on raw intelligibility, with speech form (Citation/Token) and mention (First/Second) as repeated measures, and eye-contact (With/No) and phonological reduction category (WSpoly/Nasal/d-del) as grouping variables between items. A mean intelligibility score was calculated for tokens that had appeared in more than one intelligibility experiment, rather than entering the repeated item several times in the analysis. Table 9.8 illustrates the mean intelligibility values. While citation forms are more intelligible than running speech

	Speech form		
Mention	Citation	Token	***loss***
First	.711	.541	*.17*
Second	.730	.438	*.29*
mean	*.721*	*.490*	

Table 9.8: Mean intelligibility for full set of materials (N=86), looking at first and second mentions of citation forms and running speech tokens

tokens (Form: $F_2(1, 80) = 57.1, p < .0001$), there is no main effect of eye-contact (Eye: $F_2(1, 80) = 1.06, n.s.$)[6] or phonological reduction type (Phon: $F_2(1, 80) = 1.52, n.s.$). There is a significant effect of repetition: repeated mentions of running speech tokens are less intelligible than introductory mentions, although citation forms do not differ from first to second mention (Mention: $F(1, 80) = 4.69, p < .05$; Mention x Form: $F(1, 80) = 5.53, p < .05$; Scheffé: c1=c2; c1>t1, c2>t2 ($p < .01$); t1>t2 ($p < .05$)).

In other words stressed vowel duration shows the same effect of repetition as intelligibility: repeated mentions are not only less intelligible, but also tend to have shorter stressed vowels. Correlations show intelligibility loss to be significantly related to vowel shortening in both first and second mentions (First mentions: $N = 86, r = .362, p < .001$; Second mentions: $N = 86, r = .392, p < .001$).

[6] A significant form by eye-contact interaction is shown by Scheffé to reflect a difference between citation forms that does not hold for running speech tokens: citation forms from dialogues spoken with eye-contact (mean intelligibility = .784) are more intelligible than citation forms from screened dialogues (mean intelligibility = .656). Tokens do not differ between eye-contact conditions (screened: .516, face-visible: .464), though they do differ from citation forms in both screened and face-visible conditions.

9.4 Discussion

There is little evidence to suggest that speakers shorten the duration of schwa in Weak syllable onsets of polysyllabic words when they repeat the word in spontaneous dialogue. Although schwa duration is shorter for spontaneous tokens than for matched citation form controls, schwa does not shorten further with repetition. This finding contrasts with similar analyses of stressed vowel duration, where repetition *is* associated with a reduction in duration: the stressed vowels in second mentions are shorter than in first mentions. Intelligibility is found to relate to both schwa and stressed vowel duration: intelligibility loss correlates significantly with vowel shortening, whether the vowel is stressed or metrically Weak.

When these results are compared with the results in the preceding chapters on nasal assimilation and stop deletion it becomes apparent that changes in production that occur early in the word may be more detrimental to the recognition process than changes that occur at word endings. That is not to say that word boundary reduction processes do not affect intelligibility: clearly they do, as is evidenced by the significant relation between intelligibility loss and nasal place assimilation preceding velars. Rather, because lexical competition is known to fall off rapidly as more of a word is heard (Wayland *et al.*, 1989), reduction that occurs late in a stimulus may not affect subjects' ability to recognise the target if the competition is already very low.

Where does this leave Lindblom's theory of Hyper- and Hypo-articulation? Recall that the H & H theory invokes a distinctiveness constraint to prevent speakers from economising to the point of unintelligibility: speakers economise up to but not beyond the point at which sufficient lexical contrast is maintained for successful word recognition. Clearly if lexical competition decreases as a speaker progresses through a word, then the level of hypo-articulation can afford to increase towards the ends of words. The question to be addressed, then, is what effect – if any – does the application of phonological reduction processes have on levels of lexical competition? If speakers maintain lexical distinctiveness, then they should refrain from hypo-articulating when articulatory economy leads to an increase in lexical competition. For example, consider the effects of assimilating the word-final /n/ of *been* in (9.1) and (9.2) below.

(9.1) "*The eggs have been crushed*"

(9.2) "*The eggs have been broken*"

In (9.1) the assimilated form [biŋ] is a non-word of English; the form that assimilates to the labial in (9.2), on the other hand, is identical to the English word *beam*. Thus in the first case the application of assimilation has no effect on the set of lexical competitors, while in the second the assimilation introduces a new competitor which matches the acoustic input at least as well as the target does. Lindblom's H & H theory presents speakers as cooperative dialogue agents who alter their articulatory effort according to the beliefs they hold about their listeners' needs. Do these needs include a requirement for lexical distinctiveness? If so, then speakers ought not to hypo-articulate if in doing so they introduce lexical ambiguity.

In the following chapter, therefore, the relation between phonological reduction and lexical competition is explored in greater detail. Specifically I ask whether the reduction processes of assimilation and word-final /d/-deletion ever increase the level of lexical competition, before asking, when they do, whether the increased competition reduces the likelihood of speakers applying the reduction process.

Chapter 10

Reduction and the Lexicon: Analysing Subject Responses

10.1 Introduction

In this chapter I explore the relation between the variation in production associated with phonological reduction processes such as assimilation and deletion, and the recognition of words. I argue that phonological reduction sometimes alters the size of the competitor set, with an increase in the number of competitors making the target word harder to recognise. The relation between phonological reduction and lexical competition is shown to be significant for Lindblom's H & H theory: Lindblom introduces a *distinctiveness* constraint on speaker economy whereby hypo-articulation is permitted only while lexical distinctiveness is maintained. What makes a word distinctive depends on what it has to be distinguished from. In effect, levels of hypo-articulation ought to be related to the availability of lexical items that compete with the target word. The distinctiveness constraint therefore predicts an effect of lexical competition on phonological reduction.

In the preceding three chapters I asked whether the application of phonological reduction processes contributed to the effect of repetition on intelligibility demonstrated in Chapter 6. While repeated tokens were found to be more assimilated than first mentions in one case (alveolar nasals preceding velars), no effect of repeated mention was found in several others (assimilation of alveolar nasals to following labials, word-final /d/-deletion, and schwa reduction in the metrically Weak onsets of WS polysyllables).

Because connected speech processes that effect word boundary changes (such as assimilation and stop-deletion) will frequently occur beyond a word's Uniqueness

Point (see Chapter 3 Section 3.2.5), it would seem plausible that word boundary changes are less likely to influence subjects' ability to recognise word tokens than, for example, changes in the duration and/or quality of the stressed vowel, which frequently occur before a word is uniquely identifiable. When vowel duration is examined a significant effect of repetition is found, both for the /d/-deletion materials, and for all words examined for reduction effects: in repeated mentions of landmark names, the stressed vowel is more compressed (compared with its matched citation form) than in the introductory mention. Thus reduction in stressed vowel duration mirrors the loss of intelligibility established for repeated tokens.

To test Lindblom's distinctiveness constraint on hypo-articulation we need to ascertain what effect variation in production might have on levels of lexical competition. This chapter starts, therefore, by defining the competitor set of the target word. The traditional view of lexical competition in spoken word recognition argues for incremental left-to-right matching of the input from first CV to Uniqueness Point (Marslen-Wilson and Tyler, 1980). If changes in phonological production on repetition do affect intelligibility, then a word's Uniqueness Point, for example, may well mark the domain of pragmatic (repetition) effects on articulation. In the first section I look for traditional lexical competitor effects and test the predictions against the set of candidate words offered by the intelligibility subjects: the responses of subjects who fail to recognise particular tokens correctly. I conclude that the definition of lexical competitor in terms of traditional word-initial CV cohorts and Uniqueness Points is inappropriate for the data, and, instead, propose competitor sets that are more loosely defined in terms of broad manner-based phonological classes.

Having established what counts as a competitor, I go on to explore the effects of reduction processes on recognition responses: do assimilated tokens attract more responses ending in the phonetically appropriate but lexically misleading assimilated segment, for example? If they do, then it suggests that these reductions are affecting the level of lexical competition against the target, and this may have implications for Lindblom's distinctiveness constraint. For example, where reduction processes increase rather than reduce the level of lexical competition – i.e. lead to a loss of distinctiveness – does this have a bearing on the likelihood of a process applying? In other words, do speakers refrain from assimilating or deleting pre-stress schwa or word-final /d/ when doing so makes the word token lexically ambiguous? The H & H theory would argue that hypo-articulation in repeated mentions may not be damaging for the listener because the earlier men-

tion may serve as an aid to recognition. The question, then, is whether speakers hypo-articulate introductory mentions, when there is far less contextual support to help listeners recognise the target word.

10.2 What counts as a competitor?

10.2.1 Lexical competition: a traditional approach

The likelihood that a spoken word will be recognised is, in part, a function of the word's frequency in the language and the size of the word's competitor sets: the other words in the language that sound similar to the target (see Chapter 3 Section 3.2.3). In this section I ask whether the recognition of the stimuli presented in the intelligibility experiments was affected by the size or word frequencies of the competitor set, as defined in terms of the target's word-initial cohort (Marslen-Wilson and Tyler, 1980).

For such an analysis to be informative, we need to know what the competitor set really was. Errors in subject responses from the intelligibility experiments were analysed to establish what parts of the stimulus were being matched correctly, and to locate the source of the mismatch, or recognition failure. If initial CVs were not used to key responses then explanations for recognition errors in terms of such cohorts are inappropriate.

10.2.1.1 Scoring subject responses

Each stimulus presented in an intelligibility experiment received responses from a group of listeners. For example, Table 10.1 shows the responses from the ten listeners to a citation form of the stimulus word *granite*.

Recall that only letter perfect responses were 'correct', with the total number of correct recognitions constituting the token's intelligibility. Incorrect responses were classified according to how closely they matched the target stimulus. 'Simple spelling errors' covered the inappropriate doubling of consonants, as in *grannite*, and the reduction of two consonants to one, such as *alotments*. Responses that lacked a sub-syllabic morphological affix such as the -S of *allotments* or the -ED of *abandoned* were 'affix-short', while responses with additional morphological material (invariably a plural ending) such as *shelters* or *forests* were 'extra-affix'. A category of 'phon-attempt' was introduced for those misspellings of the stimulus so extreme that it was not clear that the subject actually knew the word. Ex-

cranne it
granite
granite
granite
planet
grannite
panic
planet
granaite
granite

Table 10.1: Responses to the citation form of the stimulus word *granite*

amples like *elotments* (*allotments*), *apatchy* (*apache*), *crevis* (*crevice*) and *rivean* (*ravine*) might have been attempts at phonetic description. In this way either a **strict** ('correct') or **loose** (all classes described above) intelligibility score could be used. For the example in Table 10.1, the strict intelligibility score for *granite* is 4, while the loose score is 6.

Responses that were not obvious typos, and could not be found in either the CELEX database (Baayen *et al.*, 1995) or the Chambers English Dictionary (Schwartz *et al.*, 1988) were classed as 'non-words'; these included *bubbins, crannit, fax-them, jessoped, lagerflower, muggin* and *tellyroom* amongst others. Failure to respond at all to the stimulus (a 'null' response) was marked with '–', i.e. "Pass". For a full classification of subject responses to the stimuli used in the present study see Appendix D.

Those responses that did not fall into the above categories formed the set of real word competitors – henceforth **Real Word Responses**, or RWRs – which were analysed to assess the grounds on which subjects might be matching the stimulus input to stored lexical representations.

10.2.1.2 Locating Uniqueness Points

The results of the previous three chapters led to the suggestion that the failure to find significant effects for some word-final reduction processes was a consequence of their occurring beyond the point at which the target becomes uniquely recognisable. If a word is already uniquely identifiable before the word-final /d/ is encountered, then the phonetic quality (or indeed deletion) of the stop articulation may be irrelevant to the outcome of the lexical recognition process. To test the validity of this claim, target words were classified according to when they

become lexically unique, and correct subject responses for each category were compared. If the location of the Uniqueness Point is shown not to have an effect on successful recognition rate then UP location is unlikely to represent the domain over which the repetition effect operates.

Hypothesis 10.1 *Targets with early Uniqueness Points should be easier to recognise than targets which become unique later in the word, or even after offset: in hearing targets with early UPs subjects have the opportunity to hear more of the word after the competition has ceased, which should provide positive confirmation as to the target's identity.*

The CELEX (Baayen *et al.*, 1995) database was used to locate the conventional UP for each target stimulus. The databases for English LEMMA and WORD forms were first converted into a format that was more amenable to the use of UNIX 'grep' facilities by taking the Cobuild word frequency (for combined written and spoken frequency per 17.9 million words), orthographic form and SAM-PA transcription[1] (see Appendix B) for each word and excluding all compounds (such as *abide by*). Because there were multiple entries for certain word forms a program was run over the two databases which returned just one line per orthographic entry (provided there was no difference in transcription) with the frequencies for all previous multiple entries summed. All analyses reported below refer to these modified versions of CELEX.

Using the CELEX LEMMA database, and starting with the initial (C)(C)V word onset, the cohort size was computed for incrementally more segments through the word until either the word ended, or the cohort size fell to just 1, i.e. the target word. Thus it was possible to establish the size of the cohort as more of the word became available, along with the identity and frequency of the most frequent cohort member. Table 10.2 contains the information generated for the target stimulus *allotments*.

There is a problem in analysing responses in relation to a target's Uniqueness Point which concerns the treatment of **inflected** and **derivationally related** forms: should different inflections be counted as belonging to the same word form (in which case they share the same UP) or should they be treated as separate words (in which case the UP is often likely to be post offset)? There is evidence

[1]The CELEX database is transcribed in R.P. This poses a few problems for analyses of responses to Standard Scottish English productions of some target words, which are discussed in the text.

```
              CELEX transcription:  @.l.Q.t.m.@.n.t.
          word-initial cohort size:  946
position in cohort for 'allotment':  443
 Cobuild frequency for 'allotment':   37

                  COHORT   MOST_FREQUENT      FREQUENCY_OF_MOST
                    SIZE   COHORT_MEMBER        FREQUENT_MEMBER
search on @.         946   about  @.b.aU.t.                44684
search on @.l.        59   along  @.l.Q.N.                  5267
search on @.l.Q.       6   along  @.l.Q.N.                  5267
search on @.l.Q.t.     2   allot  @.l.Q.t.                    97
search on @.l.Q.t.m.   1   allotment  @.l.Q.t.m.@.n.t.       37
Uniqueness Point is at /m/ of allotment.

Cohort for  @.l.Q. contains:
INVERSE
 ORDER   FREQ   WORD         TRANSCRIPTION
   1        0   allopathy    @.l.Q.p.@.T.I.
   2       37   allotment    @.l.Q.t.m.@.n.t.
   3       54   aloft        @.l.Q.f.t.
   4       97   allot        @.l.Q.t.
   5      351   alongside    @.l.Q.N.s.aI.d.
   6     5267   along        @.l.Q.N.
```

Table 10.2: Information extracted from CELEX database on cohorts for *allotments* matching to V, VC, VCV, VCVC *etc.* up to Uniqueness Point. Frequency values are based on the Cobuild combined written and spoken frequency per 17.9 million words.

(Tyler *et al.*, 1993) to suggest that, in context at least, inflected forms are activated by the word stem, thus *carved* is activated with *carve*. The inflected form is fully predictable (by and large) from the root plus prior context (providing information about number, tense *etc.*). However, in isolation (as in the situation here), uniqueness can only be achieved when the affix is encountered. When presented excerpted from context, therefore, *carved* becomes unique from *carve*, *calves*, *carvery* and *carving* only when the release of the final /d/ is perceived. For this reason, the UP analysis for the set of inflected words (primarily from the /d/-deletion dataset, but also including *fallen*) was based on data from the CELEX WORD database, which lists all inflected forms separately. In this way, information is retained with respect to the relation between the possible deletion of word-final /d/ and the Uniqueness Point of the word.

Derivationally related words and compounds present a similar problem to inflected forms. Should the UP be considered post offset to distinguish *caravan* from *caravansary* and *caravanserai* (or similarly *elephant* from *elephantiasis*, and *lemon* from *lemonade*) or can it be assumed that both forms derive from a single base word *caravan*, which is uniquely identifiable from other non-*caravan*-based words by the time the /v/ at the onset of the third syllable is encountered? I decided to adopt a pragmatic approach. In many cases the derivationally related words (such as *elephantiasis* and *caravanserai*) were extremely rare, and unlikely, therefore, to influence the subject's responses. Where the derivationally related word had a word frequency of less than 5 (such as *caravanserai* and *elephantiasis*), I excluded it from the analysis. More frequent words such as *lemonade* were included and treated as competitors. The UP of *lemon* was therefore post offset, for example. Hyphenated compounds such as *green-stuff* and *swan's-down* were broken into constituent parts and assumed not to influence the location of the UP. However, words such as *greenish, greengrocer* and *swansong* were included as competitors; excluding hyphenated compounds, therefore, did not change the location of the UP for any of the stimulus words.

The set of target stimuli was classified according to whether the UP occurred:

- post offset (i.e. late) e.g. ***canoe****/canoodle*, ***pillar****/pillory*;
- on the final segment e.g. ***gazelle****/gazette*, ***camera****/camaraderie*;
- during the final syllable e.g. ***banana****/banal*, ***coconut****/coca-cola*;
- before the final syllable (i.e. early) e.g. ***abandon****/aback*, ***territory****/terrible.*

Table 10.3 shows the number of cases in each category, along with the mean intelligibility associated with the different UP locations. In the majority of cases the UP occurs in the final syllable either on or just before the final segment.

UP category	no. of polys	no. of nasals	no. of ddels	as % of total	% correct recog'd
post offset	3	4	4	13.1%	43.02%
Final segment	20	10	5	41.7%	71.66%
Final syllable	24	5	3	38.1%	70.29%
Earlier	5	0	1	7.1%	62.38%

Table 10.3: Proportion of dataset in each Uniqueness Point category, with mean intelligibility (% correct recognised)

When the UP category for each word was included as a grouping factor in a 2x2 ANOVA on mean intelligibility, with speech form (Citation/running speech Token) crossed with mention (First/Second) a significant main effect for UP location ($F_2(3, 174) = 7.20, p < .0001$) was found. A *post hoc* Scheffé test revealed that word targets with late UPs were significantly less intelligible than other targets, which did not differ from each other. In other words, while subjects predictably have difficulty recognising words which are not unique at offset, stimulus words with 'early' UPs show no advantage over words with UPs that occur nearer the word's offset.

There is a gross effect of Uniqueness Point, then, with words that are conventionally unique on or before the final segment being more intelligible than words that still have traditional lexical competitors at offset. There is, however, **no fine distinction to be made between words that become unique extremely early, and those that become unique only just before offset**. The intelligibility of these target words is the same. The location of traditional Uniqueness Points, then, appears not to differentiate the target stimuli in this set of data, beyond distinguishing pre and post offset.

The theory underlying the Uniqueness Point of a word assumes that subjects match the input incrementally from left to right as more phonetic information about the word's identity is made available. But given the level of hypoarticulation that Lindblom's H & H theory predicts in the context of communicative discourse, perhaps this presents a rather optimistic account of the acoustic information available to listeners? Is there always sufficient information in the acoustic stream to enable listeners to access the correct word-initial cohort?

10.2.1.3 Do responses fall within the same word-initial cohort?

According to the traditional Cohort model of word recognition, the set of RWRs should be dominated by competitors from the target word's CV cohort that are more frequent in the language than the target itself. To test this prediction, I examined how many RWRs fall within the same word-initial cohort, i.e. match to #(C)(C)V, and compared their relative word frequencies.

The CELEX LEMMA[2] database was used to construct word-initial cohorts for each stimulus word form, based on the #(C)(C)V structure of the target word. Subject responses were then examined to see whether the most frequently offered competitor to the stimulus word was

- a member of the same cohort
- more frequent than the stimulus word.

For example, given the set of responses to the word *bakery* in Table 10.4, a list was created, ordered by the most frequently offered response, and transcribed according to the SAM-PA transcription in the CELEX database (see Appendix B), resulting in the output in Table 10.5.

TARGET	DUR(ms)	RESPONSES
bakery	367.94	1 1 1 1 - 1 1 1 1
bakery	379.50	good baby 1 1 1 1 -hey 1 -
bakery	435.12	1 1 1 1 1 1 1 1 1
bakery	408.25	1 1 1 1 1 1 1 1 1
bakery	367.94	1 1 beggary 1 1 1 1 1 1
bakery	387.75	1 1 1 1 1 1 1 1 1
bakery	291.44	baking bacon bacon 1 1 baby bacon 1 teacake
bakery	283.12	1 1 1 bacon bacon -group 1 1 bacon
bakery	435.12	1 1 1 1 1 1 1 1 1
bakery	428.94	1 1 1 1 1 1 1 bakerey 1

Table 10.4: Responses to stimulus word *bakery*

For the example in Table 10.5, the most frequent response is clearly the stimulus itself, *bakery*. Real Word Responses (RWRs) that do not match the stimulus include *bacon, baby* and *baking* all of which fall within the same word-initial CV cohort. The first twenty members of the /#beɪ/ cohort are presented in

[2]Cohort analyses were conducted on both the LEMMA and WORD databases. Since the analyses conform on the critical results only the details of the LEMMA based searches are reported; the WORD database is large and unwieldy, and tends to obscure trends.

NO.	RESPONSE	TRANSCRIPTION	RESPONSE-TYPE
73	1		Correct
6	bacon	b.eI.k.@.n.	RWR
2	baby	b.eI.b.I.	RWR
2	-		Null Response
1	teacake	t.i:.k.eI.k.	RWR
1	good	g.U.d.	RWR
1	beggary	b.E.g.@.r.I.	RWR
1	baking	b.eI.k.I.N.	RWR
1	bakerey		Typo
1	-hey		Non Word
1	-group		Non Word

Table 10.5: Responses to stimulus word *bakery*, ordered by the most frequently offered responses, and transcribed using the CELEX SAM-PA transcription

Table 10.6, where it can be seen that the competitors *bacon* and *baby* are more frequent cohort members than the target word *bakery*.

Compare this with the set of responses to the stimulus word *wagon* in Table 10.7, which, when listed, looks like Table 10.8. Again the most dominant response is the correct recognition of *wagon*, but the next most likely response – the dominant RWR – is *dragon*, which, though a 'rhyme prime' (Marslen-Wilson and Zwitserlood, 1989; Marslen-Wilson *et al.*, 1996), does not share the same word-initial cohort as the target stimulus. Indeed, not one of the incorrect responses shared the same cohort as the target.

The question, then, is how many of the target stimuli received RWRs that fell within the same word-initial cohort, and were these cohort responses more frequent than the stimulus word? In other words, did the majority of target stimuli receive responses like those to *bakery* or like those to *wagon*?

The total dataset of 84 items comprised 13 different word forms which may undergo word-final /d/-deletion, 19 word forms which may undergo word-final place assimilation, and 52 polysyllabic words, of which 17 were WS and 35 were SW. Although the full set of polysyllabic words contained 59 items (see Chapter 9), 7 of these were included within the /d/-deletion and nasal assimilation datasets, so when the data was pooled the maximum number of unique word forms was 84. Of these 84 word forms, 4 could not be included since they failed to elicit any Real Word Responses (*abandoned, crocodiles, gazelles* and *rocket*), and one word, *desert*, was excluded because the only RWRs – six responses of *dessert* – were most probably 'typos' (given the poor standard of spelling evident in the

Position in Cohort	Cobuild Frequency	Stimulus Word	Transcription
1	4622	**baby**	b.eI.b.I.
2	1931	basic	b.eI.s.I.k.
3	1499	basis	b.eI.s.I.s.
4	1462	base	b.eI.s.
5	829	bait	b.eI.t.
6	643	bay	b.eI.
7	526	basically	b.eI.s.I.k.@.l.I.
8	423	bake	b.eI.k.
9	341	basin	b.eI.s.n,.
10	311	baker	b.eI.k.@.r*.
11	288	**bacon**	b.eI.k.@.n.
12	236	basement	b.eI.s.m.@.n.t.
13	209	bail	b.eI.l.
14	168	bathe	b.eI.D.
15	128	BA	b.eI. (*sic*)
16	110	baseball	b.eI.s.b.O:.l.
17	88	bathing	b.eI.D.I.N.
18	83	bass	b.eI.s.
19	83	**bakery**	b.eI.k.@.r.I.
20	66	bayonet	b.eI.@.n.I.t.

Table 10.6: First twenty members of the word-initial cohort for /#beɪ/. Frequency taken from Cobuild corpus of 17.9 million words.

TARGET	DUR(ms)	RESPONSES
wagon	325.00	- laddie 1 1 1 1 1 waddy laddie
wagon	328.69	dragon 1 1 dragon 1 1 1 1 1
wagon	364.00	1 dragon 1 dragon 1 ragging 1 dragon 1
wagon	325.00	1 - - 1 1 slagheap laddie dragon worry
wagon	299.19	1 1 1 1 1 1* 1 1 1
wagon	364.00	dragon 1 1 dragon 1 1 1* 1 1

Table 10.7: Responses to stimulus word *wagon*

NO.	RESPONSE	TRANSCRIPTION	RESPONSE-TYPE
34	1		Correct
8	dragon	d.r.{.g.@.n.	RWR
3	laddie	l.{.d.I.	RWR
3	-		Null Response
2	1*		Simple Spelling Error: e.g.\ waggon
1	worry	w.V.r.I.	RWR
1	waddy		Non Word
1	slagheap	s.l.{.g.h.i:.p.	RWR
1	ragging	r.{.g.I.N.	RWR

Table 10.8: Responses to stimulus word *wagon*, ordered by the most frequently offered responses, and transcribed using the CELEX SAM-PA transcription

intelligibility experiments as a whole).

Of the 79 stimuli that could be analysed, roughly 60% of the target words elicited at least one response within the same word-initial cohort (48/79 = 61%). Of these, just over half (26/48 or 54%) elicited one particular word-initial cohort member more often than any other response. For example, the RWR that occurred most often to the stimulus *carved* was *card*, and to *totem* was *total* (see Table 10.9).

An examination of the relative positions within the word-initial cohort of the target stimulus and the dominant cohort response revealed that of the 24 cases that could be analysed by far the majority (18/24 or 75%) involved cohort member responses that had a higher lexical frequency than the target word. In other words, when subjects match to CV onset they are likely to offer word candidates which occur frequently in the language. However, **fewer than a quarter of all target stimuli (18/79 or 23%) elicited a Real Word Response pattern that was dominated by a more highly frequent member of the target's word-initial cohort**. Clearly most response profiles involved dominant word candidates that matched less strictly to the target stimulus.

A total of 31 stimulus words (39%) received RWRs that **fell entirely outwith the word-initial cohort** (i.e. that had response profiles similar to that for *wagon*). For twelve of these target words (38.7%), the total number of Real Word Responses was five or fewer. Such small numbers of RWRs are associated with the more intelligible word forms, the words which were easily recognised. It may be that the few incorrect responses to these words should be treated as potentially suspect. But even disregarding the stimuli for which there were five or fewer RWRs, slightly less than one third of the stimulus set (30.6%) elicited responses that entirely failed to access the correct word-initial cohort. Not one of

Target stimulus			Most dominant RWR		
Word	Cobuild freq	Cohort position	Word	Cobuild freq	Cohort position
bakery	83	19th	**bacon**	288	11th
blacksmith	48	13th	**black**	6545	1st
canoe	109	161st	**career**	1285	18th
caravan	179	45th	**cabin**	531	25th
carved	350	9th	**card**	1259	3rd
cattle	568	21st	**catholic**	747	16th
cavalry	116	54th	calvary	2	282nd
crossing	177	3rd	**cross**	1973	1st
diamond†	140	18th	diving	95	22nd
disused	30	422nd	**dishes**	237	128th
fallen†	848	13th	**falling**	932	12th
farmed†	11	45th	**farmer**	562	9th
gold	1575	2nd	**go**	52264	1st
level	4588	2nd	leopard	151	26th
lion	444	8th	**line**	5665	5th
overgrown	54	59th	overgrowing	*not available**	
overnight	332	21st	overly	30	87th
pillar	285	14th	**pillow**	350	12th
popular	1440	7th	popgroup	*not available**	
savannah	42	88th	**survival**	677	17th
settlement	471	35th	sentiment	213	54th
telephone	2241	5th	television	2043	6th
totem	21	17th	**total**	2520	1st
tribal	199	7th	**triangle**	219	5th
waterfall	137	16th	**water**	8418	1st
waterhole	18	44th	**water**	8418	1st

Table 10.9: Word frequency (Cobuild 17.9 million corpus) and cohort position of stimulus word and most dominant Real Word Response for all cases where most dominant RWR falls within same word-initial cohort. The more frequent cohort member is in bold.
(†: use of WORD rather than LEMMA dictionary)
(*: these words do not appear in the CELEX database)

the 30 responses to the target *beeches*, for example, started with the appropriate CV onset /#bi/: the most dominant response, *peaches* (N=15), fails to match at onset.

The metrical structure of the target stimulus has a significant effect on the kind of matches elicited, with a high proportion of WS polysyllables failing to elicit word-initial cohort matches (see Table 10.10). Targets such as *canal* and *collapsed* are less likely to elicit word-initial cohort matches than SW targets such as *cattle* and *crossing* ($\chi^2 = 4.684, p < .05$). Indeed, fewer WS words managed to elicit any kind of cohort match than failed outright.

Pattern of RWRs	**Metrical Structure of Stimulus**	
	WS	SW
at least one response in same word-initial cohort	7	41
not one response in same word-initial cohort	11	20

Table 10.10: Effect of metrical structure on likelihood of eliciting Real Word Responses in the same word-initial cohort

Clearly, then, the relation between failed recognition and the size and frequency of the competitor set matched to word onset is not a simple one: subjects did not consistently respond to target stimuli with more frequent members of the target's word-initial cohort. This fact is reinforced by the failure to find any significant correlation between intelligibility, word-initial cohort size and word frequency (in all cases $R < 0.1$).

10.2.2 Lexical competition: matching to length and rhythm

Given the failure of many Real Word Responses to derive from the same word-initial cohort as the stimulus word, is it possible to define lexical competition in a way that captures a larger proportion of RWRs? What are subjects matching to if not standard CV onsets?

It has been suggested (see Kelly, 1993, for a full discussion) that listeners use the metrical structure of words to segment the acoustic stream, positing word onsets at Strong syllables. If subjects are sensitive to metrical structure in this way, then perhaps their responses will match the metrical structure of the target stimulus, even if they fail to match to CV onset.

Traditional CV cohort analyses also fail to take account of word length. Clearly short stimuli are unlikely to elicit polysyllabic responses that match at CV onset but are significantly longer than the target. It may be that subjects are at least as sensitive to stimulus length as to the phonetic content of the stimulus.

To what extent, then, might subjects be matching to the number of syllables or the metrical structure of the target stimulus, and did they favour one over the other?

The polysyllabic stimulus set (N=55) was classified according to whether the initial syllable was metrically Weak (i.e. contained a reduced vowel e.g. /ə/ or /ɪ/) or metrically Strong (i.e. contained a full vowel) (Cutler and Norris, 1988). Real Word Responses were then examined to see whether they shared the same number of syllables as the target stimulus, and/or whether they shared the same metrical structure; that is, whether WS polysyllables elicited RWRs that had Weak first syllables followed by Strong second syllables, and SW polysyllables elicited RWRs that had Strong first but Weak second syllables.

Target words were then grouped according to whether:

1. more RWRs matched correctly to metrical structure than syllable number (sylls < metr);
2. the number of matches to metrical structure and syllable number was the same (sylls = metr);
3. more RWRs matched correctly to the number of syllables than to the metrical structure (sylls > metr).

The distribution of responses in each category type can be found in Table 10.11.

			Metrical Structure:		**Syllable Number:**	
Category	**N**	**%**	WS	SW	Three	Two
1. sylls < metr	17	31%	1	16	12	5
2. sylls = metr	23	42%	6	17	8	15
3. sylls > metr	15	27%	10	5	1	14

Table 10.11: Distribution of Real Word Responses that match on syllable number and/or metrical structure for SW and WS initial words with two or three syllables

In general, Weak-initial words elicited RWRs that matched to syllable number rather than metrical structure, while polysyllabic words of three or more syllables matched at least as well to metrical structure as syllable length.

Let us suppose that subjects tend to respond to polysyllabic stimuli with a SW bisyllabic word, since the lexicon offers many more SW than WS words, and more words of two syllables than of any other length[3]. Even if subjects know little about the words presented, but respond randomly, then:

- for SW stimuli with 2 syllables the responses will tend to be matched correctly on both metrical structure and syllable length;
- for SW stimuli of 3+ syllable length the metrical structure will be appropriate but the number of syllables wrong (= category 1 above);
- for WS stimuli with 2 syllables the metrical structure will be wrong but the number of syllables will match correctly (= category 3 above);
- for WS stimuli of 3+ syllable length a SW bisyllabic response will fail to match either metrical structure or number of syllables.

The question is whether the preference observed above, for WS words to match to syllable number, and Strong-initial words of three or more syllables to match better to metrical structure, can simply be accounted for entirely by a tendency for listeners to offer SW bisyllables as responses to all polysyllabic stimuli. The answer appears to be negative: responses differ according to the metrical structure of the word (see Table 10.12). Responses to WS stimuli are predominantly other than bisyllabic SW words, while there is no consistent response to SW stimuli. The difference is significant ($\chi^2 = 12.51, p < .001$).

Predominant response	**Metrical Structure of Stimulus**	
	WS	SW
2 sylls(SW) > Other	1	20
2 sylls(SW) < Other	16	15

Table 10.12: Preference for SW bisyllabic responses to stimuli with Weak- and Strong-initial syllables

The SW stimulus set was then split into words with 2 and words with 3 or more syllables, and the preference for bisyllabic SW responses examined (see Table 10.13). It was hypothesised that SW bisyllabic stimuli would elicit a relatively

[3]CELEX contains 8,259 unique monosyllabic words, 32,215 bisyllabic words, 28,598 three-syllable words, 14,378 four-syllable words, and roughly 6,000 words of more than four syllables in length.

high proportion of SW bisyllabic responses, while SW trisyllabic stimuli would elicit a greater number of 'Other' responses. However, there was no significant difference in the distribution of RWRs ($\chi^2 < 0.1, n.s.$).

SW	**Predominant response**	
stimulus	2 syll(SW)	Other
Two sylls	13	9
Three sylls	7	6

Table 10.13: Preference for SW bisyllabic responses to SW stimuli with two or three syllables

A close inspection of the RWRs reveals that SW bisyllables receive a number of monosyllabic responses, while SW trisyllables receive a large number of bisyllabic responses. In other words, responses reflect the potential loss of any Weak syllable. Subjects appear to match the stimulus input on to representations which are one syllable shorter than the fully articulated form of the stimulus. The difficulty in recognising the onset of WS polysyllables, then, reflects a more general problem of recognising Weak syllables.

10.2.3 Lexical competition: defining 'loose' cohorts

The results above show that subjects frequently failed to match the incoming stimulus to the correct word-initial cohort, the **strict** CV cohorts advocated by Marslen-Wilson and colleagues. The question remains whether there is a pattern to the responses that could be captured by defining a **looser** cohort structure. Is it possible to define a less fully determined phonologically based cohort which covers at least a sizeable majority of the responses?

It has long been known that some phonological categories are more confusable than others (Miller and Nicely, 1955): while voicing and nasality appear to be resistant to the effects of noise, for example, information about place of articulation is much more vulnerable.

A perusal of the error responses made in the intelligibility experiments suggests that the confusions experienced by subjects frequently related to place categories; subjects responded to *crane* with *train*, for example. Segments were grouped according to the manner-based classes in Table 10.14, and subject responses re-examined to see how many RWRs matched at least to the broad segment class rather than the uniquely specified CV segments.

Class	Members
Stops	/p b t d k g/
Fricatives	/f v θ ð s z ʃ ʒ/
Nasals	/m n ŋ/
Semi-Vowels	/l r w j/
Tense Vowels	/i ɑ ɔ u ɜ aɪ eɪ oɪ aʊ oʊ ɪə uə/
Lax Vowels	/ɪ ɛ a ʌ ɒ ʊ/

Table 10.14: Broad manner-based classification of segments used for loose cohort matching

Of course matching *crane* (/kreɪn/) to the loose CCVC cohort structure in (10.1) would capture a huge number of competitors (N=337, based on a CELEX search with regular expressions).

(10.1) [stop][rlwj][V_{tense}][nasal]

So for each stimulus word, **a comparison was made between the number of RWRs which matched strictly to the consonantal onset but loosely to the vowel, and the number of RWRs matching strictly to the vowel but only loosely to the consonantal onset.**

carved

k. A:. v. d. N = 147

Loose Cohort		RWRs
k	A:	103
stop	A:	104
k	V_{tense}	132
stop	V_{tense}	135
k	V_{any}	138
stop	V_{any}	142
remainder		5

Table 10.15: Example illustrating relative size of different 'loose' matches to the stimulus *carved*

For the example in Table 10.15, loosening the match criterion for the initial /k/ segment (i.e. matching to [stop] rather than /k/) fails to buy any substantial improvement in the number of RWRs that get 'captured' by the cohort structure (103 → 104). On the other hand, extending the vowel description from /ɑ/ to [V_{tense}] improves the number of RWR onset matches from 103 to 132. This is because several subjects responded with words like *curve, curved* and *curse.* For

this particular stimulus, then, a word-initial cohort based on a match to CV_{tense} would incorporate a substantial majority of the subject responses.

disused

d. I. s. j. u:. z. d. N = 41

Loose Cohort		RWRs
d	I	28
stop	I	38
d	V_{lax}	28
stop	V_{lax}	38
d	V_{any}	28
stop	V_{any}	38
remainder		3

Table 10.16: Example illustrating relative size of different 'loose' matches to the stimulus *disused*

The data in Table 10.16 presents a different picture. Here, loosening the match criterion for the vowel has no effect on the number of RWRs that get captured by the cohort description, whereas a loose cohort that matches to a [stop]/ɪ/ onset will include an additional 10 responses (9 responses of *tissues* and one of *tissue*).

Clearly, the responses to each target stimulus in the dataset can be analysed in this way. But what factors dictate whether a particular target is likely to match better to a CV_{loose} frame than a $C_{loose}V$ frame? Can any kind of pattern be found that would enable us to predict which targets would fall into which loose cohort category?

Loose cohort analyses of the sort illustrated in Tables 10.15 and 10.16 were performed on the full set of stimulus words (N=79). In light of the RWRs offered to the metrically Weak-initial targets, it was decided to exclude the whole WS word set from the subsequent analyses presented here. The weak onset invariably resulted in subjects failing to match to the weak vowel, that is, to words sharing a $C_{loose}V$ frame, and, while several stimulus words (N=7) elicited RWRs that shared the same consonantal onset (i.e. fell within a CV_{loose} frame), the remaining 11 items could not be easily classified. In particular, the responses to *giraffe* and *collapsed* suggest that some words may undergo a process of resyllabification with the initial schwa being deleted. The loose cohort predictions for these target types is rather different from those presented here, and are discussed more fully in Section 10.3.2.1. Since one of these WS words (*saloon*) was also a member of the nasal assimilation set, and one (*submerged*) a member of the /d/-deletion set, the final number of items analysed for loose cohorts was 61

(12[ddel]+18[nasal]+31[SW]).

All responses to the 61 stimulus words were analysed in terms of matches to loose cohorts of the form CV_{loose} and $C_{loose}V$, and each stimulus classified according to which cohort frame captured the most RWRs (see Table 10.17).

Loose Cohort Category	**N**	**Stimulus Words**
traditional C V cohort	9	*bakery, blacksmith, caravan, cavalry, overgrown, telephone, totem, waterfall, waterhole*
C V_{loose}	21	*camera, carved, cattle, cobbled, coconut, crevice, crossing, elephant, fallen, farmed, Indian, iron, lemon, monastery, overnight, pillars, pine, popular, settlement, seven, swan*
C_{loose} V	19	*bandit, beeches, chapel, crane, diamond, disused, gold, golden, granite, green, lion, pelicans, picket, round, Saxon, shelter, train, tribal, wagon*
match to something post onset	7	*forest, monument, pebbled, poisoned, roman, village, walled*
others	3	*level, limestone, tourist*
no good match (possibly too short)	2	*old, owned*

Table 10.17: Distribution of target stimuli to varying loose cohort frames

The majority of target stimuli (56/61) fell within four 'cohort' categories, with RWRs matching best to:

- C V_{loose} cohorts;
- $C_{loose}V$ cohorts;
- traditional CV cohorts;
- some later part of the word stimulus.

I deal with the five exceptional/problem cases first, to eliminate them from further consideration.

The intelligibility of the targets *old* and *owned* was extremely low, with subjects offering monosyllables of varying degrees of match. This was primarily due to the very short duration of all tokens of these two word forms (mean duration of *owned* = 275.3 ms, or 229.8 ms when the longer, more intelligible citation forms are excluded; mean duration of *old* = 224.8 ms, or 170.1 ms for running speech tokens only). Given there was so little acoustic information available to the subjects it was decided to exclude these two cases.

The pattern of responses to the target *level* was too variable to allocate this word easily to any of the four cohort categories. If the incorrectly spelled 'leapoard' is corrected to *leopard*, then *leopard* becomes the most dominant RWR (5/13), which would suggest placing *level* in the traditional CV cohort category. The second most dominant response (4/13) is *river*; this can only be captured by a $C_{loose}V_{loose}$ cohort, which has not been included on the grounds that it is too broad a match to be a useful predictor of likely responses[4].

There were only three incorrect responses to the target *limestone*, yielding the RWRs *oyster, oysters* and *plimsoll.* The location of the match between the target word and these alternative candidates is not immediately obvious, and given so few responses it was decided to leave this word aside as an exception.

The stimulus *tourist* was problematic in that of the 7 Real Word Responses, 6 started with the onset /ʧ/ (*cheers(×2), choice(×2), cherries, church*) and the other with /ʤ/ (*jury*). Whilst it is not surprising that /tʊə/ may be perceived as [ʧV], especially when overlayed with noise, it seems inappropriate to classify these responses as an exact match to the consonantal onset, and they clearly fail to match to the vowel.

10.2.3.1 Matches to 'loose' cohorts

The two 'loose' categories – CV_{loose} and $C_{loose}V$ – attracted two-thirds of the dataset. Whether a word falls within one category or the other appears to depend largely on the phonetic characteristics of the onset segment, as Table 10.18 makes evident.

As a general rule, **stimulus words starting with voiceless stops or fricatives activate response words that match to the exact consonant but only**

[4]The best frame for matching RWRs to *level* is actually: [l/r][ɪ/ɛ][C_{labial}][$syll_{Weak}$].

Loose Cohort Category			
[C V_{loose}]		[C_{loose} V]	
	STOPS/AFFRICATES		
#k—	*camera*	#g—	*gold*
	carved		*golden*
	cattle	#d—	*diamond*
	cobbled		*disused*
	coconut	#b—	*bandit*
			beeches
#p—	*pine*	#p—	*pelicans*
	pillars		*picket*
	popular		
		#ʧ—	*chapel*
	FRICATIVES		
#f—	*fallen*		
	farmed		
#s—	*settlement*	#s—	*Saxon*
	seven		
	swan		
		#ʃ—	*shelter*
	LIQUIDS AND GLIDES		
#l—	*lemon*	#l—	*lion*
		#r—	*round*
		#w—	*wagon*
	CC CLUSTER		
#[stop]r—	*crevice*	#[stop]r—	*crane*
	crossing		*green*
			granite
			train
			tribal
	NASALS		
#m—	*monument*		
	VOWEL ONSET		
#V	*elephant*		
	Indian		
	iron		
	overnight		

Table 10.18: The relation between loose cohort category and phonological class of onset segment

loosely to the vowel. This observation holds true for /k/-initial words (*camera, carved, cattle, cobbled, coconut, crevice* and *crossing*), and for words starting with /f/ (*fallen* and *farmed*).

There are, of course, a few exceptions to the general principle, namely the /p/-initial words *picket* and *pelicans*, and the words *Saxon, shelter* and *chapel.* These stimuli would be predicted to fall within the CV_{loose} cohort category but in fact are matched by RWRs that best fit the $C_{loose}V$ frame. Of these five exceptions, however, only two (*picket* and *Saxon*) were matched to more than three RWRs. The fact that *pelicans, shelter* and *chapel* received so few incorrect responses means that their classification should be regarded with suspicion. Responses to the word *picket* (freq=154) were dominated by the more frequent *ticket* (freq=659) while responses to *Saxon* (freq=49) were dominated by the words *taxi* (freq=645) and *factory* (freq=1064). Clearly, where there are one or two highly frequent competitors which match closely to the target word, these will bias the loose cohort matching category in their favour[5].

It should be borne in mind that the noise overlay used in these intelligibility studies was designed to distort the signal in proportion to the signal amplitude; this may have resulted in 'noisy' parts of the signal – fricatives and aspiration, for example – being overly emphasised, leading subjects to respond with words sharing similar noisy onsets; if responses match to the first part of the signal which is perceptually dominant, then RWRs to words that start with significant high frequency energy spectra (i.e. voiceless stops or fricatives) will tend to match to the initial consonant, thereby placing them in the CV_{loose} category.

Stimulus words that start with voiced stops, or consonant clusters, on the other hand, tend to match well to the stressed vowel, but only loosely to the features of the initial consonant. (Thus responses to *diamond* included *timing* and *tyrant*, responses to *gold* included *cone, told* and *boulder*, and the most dominant RWR to *beeches* was *peaches.*) Cluster onsets were frequently responded to with similar clusters or with affricates, so that *granite* received responses of *planet, green* was heard as *dream, jean* and *cream*, whilst the most dominant response to *crane* was *train.* There are two exceptions where [stop]/r/ clusters appear to match better to CV_{loose} cohorts than to $C_{loose}V$ cohorts: subjects responded with *crevasse* and *cruise* to *crevice*, and *crust* and *crushing* to *crossing.* However, the pattern of responses to *crossing*, in particular, is difficult to classify: there are several responses that match to the vowel but

[5]Note also, that /p/ is not an 'acoustically robust' segment (Hawkins and Warren, 1994, page 501).

not to a loose match at consonant onset, such as the RWRs *fossil, possum* and *possibly.* Because these words fail to match to a [stop][rlwj] onset, the dominant response pattern could not be categorised as $C_{loose}V$, and consequently *crust* and *crushing* responses result in *crossing* being classified as CV_{loose}.

The responses to stimulus words starting with laterals and glides more frequently fall into the $C_{loose}V$ category, but there are few examples, and there is conflicting data with respect to lateral onsets. For example, while *lemon* elicits /l/-initial responses such as *limb* and *linen*, the stimulus *lion* attracts responses to the vowel, as in *wine, white* and *bile*, and *limestone* fails to match either to the consonant or the vowel.

Words with vowel onsets tend to activate RWRs which, though they start with vowels, do not always match the precise identity of the vowel, i.e. they match to CV_{loose} cohorts. For example, *elephants* elicits the responses *athletes* and *airfleet*, while the dominant RWR to *Indian* is *ending.*

There is one further observation worth consideration. For the dataset used in this analysis, there appears to be a tendency for the words that match to CV_{loose} cohorts to have lax vowels, while the words that match to $C_{loose}V$ cohorts have tense vowels, especially diphthongs. (For example, words that match to CV_{loose} cohorts include *camera, cattle, seven, lemon, popular, monument* and *Indian*, whereas the set of words that match to $C_{loose}V$ includes *train, crane, diamond, tribal, beeches, golden* and *round.*) The exceptions to this general trend are words with lax vowels that match better to a loose cohort based on a strict vowel match (such as *bandit, disused, wagon* and *granite*).

The results of a 2x2 Chi-square (see Table 10.19) shows a significant difference in the distribution of tense and lax vowels to the two CV_{loose} and $C_{loose}V$ cohort categories ($\chi^2 = 4.95, p < .05$)[6]. If the problem cases discussed above (*pelicans, picket, shelter* and *chapel*) are excluded, then the proportion of tense to lax vowels in the $C_{loose}V$ cohort category becomes 10:5, and the value for χ^2 rises to 8.35,

[6]It should be noted that the word *fallen* was classified here as a 'lax' vowel. The Scottish vowel system does not distinguish /ɒ/ and /ɔ/ although there may be a length difference between the vowels in *cot* and *caught.* The primary feature which distinguishes tense from lax vowels is, of course, syllable structure: while tense vowels can occur in open syllables, lax vowels can occur only in closed syllables. But, as Ladefoged (1982) observes (p80 ff.), the vowel /ɔ/ is somewhat unusual in that for accents of English that make no /ɒ/-/ɔ/ distinction, /ɔ/ patterns with lax vowels in terms of the kinds of syllables in which it occurs. For example, only lax vowels – and /ɔ/ – occur in syllables closed by /ŋ/, such as the word *long.* Given this fact, and the observation that the different productions of *fallen* sounded as if they were produced with a vowel much more like RP /ɒ/ than RP /ɔ/, it was decided to classify these vowels as lax for current purposes.

which is significant at the $p < .01$ level.

Loose Cohort Category	Vowel Type LAX	TENSE
C V_{loose}	17	4
C_{loose} V	9	10

Table 10.19: Number of stimulus words with RWRs that match best to one of two loose cohort categories (CV_{loose} and $C_{loose}V$) grouped according to whether the stressed vowel is tense or lax

As a general rule, then, **words that elicit responses that match to a CV_{loose} cohort tend to start with voiceless stops and fricatives, and contain lax vowels, while words that elicit responses that match to a $C_{loose}V$ cohort tend to start with less acoustically dominating consonants (voiced stops, liquids, glides) but tense vowels.**

The CELEX database was searched to ascertain whether this was a general pattern that held for the distributions of CV onsets in English: can words be divided into those that have strident[7], easy to recognise consonants combined with lax vowels, and less clear consonants with tense vowels? The number of words that started with various consonant classes combined with either lax or tense vowels was summed and the relative proportions compared. The consonant onset groups were /ptk/, /bdg/, /mnŋ/, /lrwj/, /fsʃθ/, /vzʒð/, and [stop]/r/ clusters. The pattern observed for the subset of data described above was not replicated for the lexicon as a whole: the proportion of words with tense to lax vowels for each consonant onset group was more or less constant: roughly 55% of words have lax vowels, 45% of words have tense vowels and 5% of words start with schwa. In no consonant onset category are there more words with tense than lax vowels.

However, the CELEX search of the whole corpus did not consider word frequency or length and it may be that the overall balance of tense/lax vowel to acoustically 'strong' or 'weak' consonant onset observed above is affected by large numbers of long, low frequency words which are not likely to appear as Real Word Responses in a lexical identification task. If these were eliminated it is possible that the distribution of vowel-type to consonant-type onsets in CELEX would look more like the set of Real Word Response data presented here.

[7] I use the term 'strident' here in a non-standard sense. Although stop consonants like /p,t,k/ are not classified as +strident in standard SPE-like feature descriptions, the burst/aspiration phase of voiceless stops (especially /t/ and /k/) is acoustically similar to the high frequency energy spectra of the voiceless fricatives /s/ and /ʃ/. It is this similarity that I am trying to capture in using the term 'strident' here.

10.2.3.2 Matches to traditional CV cohorts

Some of the target stimuli did not match to either a CV_{loose} or $C_{loose}V$ frame. There were nine stimulus words that matched to traditional CV onsets. Three of these – *blacksmith, waterfall* and *waterhole* – elicited responses to the more frequent word contained within them (*black* and *water*) and one, *overgrown*, was responded to primarily with alternative *over-* words. A further three words (*telephone, caravan* and *cavalry*) received few (N $\leq$ 3) incorrect responses, suggesting that these polysyllabic words were essentially 'too intelligible' for a loose cohort analysis to be appropriate. The remaining two words in this set are *totem* and *bakery.* RWRs to both these stimuli were dominated by more frequent members of the same cohort: *total* for the rare word *totem*, and *baby/bacon* for *bakery.*

10.2.3.3 RWRs which match post onset

A group of seven stimulus words could not be categorised according to any kind of cohort match from word onset. The dominant RWRs to this set of targets appear, rather, to match to a later part of the word, most frequently the coda of the final syllable. So, for example, 8 of the 9 RWRs to *village* were words ending in /ɪʤ/: *courage, forage, image, ridge, porridge, sewage* and *voyage*(×2). Similarly, responses to *forest* included *tourist, provost, furthest* and *first*, while RWRs to *walled* included *world, bald* and *old.* The most dominant response to *pebbled* was *table* (N=29/59). Of the 59 RWRs to *pebbled*, a total of 45 matched to the sequence[V_{any}bl̩].

All post onset matching targets except one (*walled*) were polysyllabic, and in all but two cases the match was to a word-final consonant cluster involving at least one alveolar (/st/, /nt/, /bl̩d/, /z̩nd/, /ld/). The two exceptions were *village* (matched to an affricate) and *roman*, which is a bit of an oddity, being matched primarily on nasality (/n/ and/or /m/) and to the /mən/ offset. (For example, RWRs to *roman* included *women*(8), *woman*(6), *moment*(4) and *bomb*(6).)

10.2.4 Summary

An analysis of subject responses (RWRs) reveals that failure to recognise a word token presented in an intelligibility experiment is not simply a result of subjects matching to a more frequent word-initial cohort member: roughly one third of the target stimuli are found to elicit RWRs which fall entirely outwith the word-initial cohort, while only one quarter of the dataset elicited a RWR pattern that was

dominated by a more highly frequent member of the target's word-initial cohort.

But although subjects frequently fail to match the incoming stimulus to the strict word-initial #(C)(C)V cohorts advocated by Marslen-Wilson and his colleagues, it is possible to define a 'loose' cohort structure which captures a sizeable majority of subject responses. Target stimuli tend to match to CV_{loose} or $C_{loose}V$ competitor cohorts, where 'loose' refers to a broad manner-based class of segments, and where membership of each category is dependent upon the phonetic characteristics of the word onset. Real Word Responses that match exactly to the consonant onset but only loosely to the vowel are activated by stimulus words starting with voiceless stops or fricatives; stimulus words that start with voiced stops, or consonant clusters, on the other hand, match well to the stressed vowel, but only loosely to the features of the consonantal onset. For the data in this study, there is also a significant correspondence with vowel type: stimulus words with lax vowels tend to elicit CV_{loose} cohorts, while stimuli with tense vowels elicit $C_{loose}V$ cohorts.

The metrical structure of the stimulus word is also shown to have a significant effect on the kind of matches elicited: words with metrically Weak initial syllables, such as *canal*, are less likely to elicit word-initial cohort matches than words with Strong initial syllables, such as *cattle*.

It was also observed that responses to polysyllabic targets are frequently one syllable shorter than the fully articulated form of the stimulus word. This finding suggests that the reduction and possible deletion of vowels in metrically Weak syllables may have disruptive effects on the successful recognition of words in isolation.

Having now defined lexical competition in a manner that accounts for the sort of candidate words offered as responses to our degraded natural speech stimuli, we can return to the primary concern of this chapter, which is the relation between lexical competition and phonological reduction. Lindblom argues for a distinctiveness constraint whereby hypo-articulation is restricted by a listener requirement for lexical discriminability: if speakers hypo-articulate beyond the point of lexical distinctiveness then the activation of a competing candidate may result in the listener failing to recognise the target word. In the following section I ask whether the phonological reduction processes examined in earlier chapters ever lead to lexical ambiguity, or at least a change in lexical competitor set that would result in a close competitor challenging the target for recognition. Evidence of such a change would have implications for Lindblom's H & H theory, which could

then be tested.

10.3 Does phonological reduction affect lexical competition?

In the previous section I argued for a model of competitor cohorts that is based on a loose match between input and target, where cohorts share some but not all of the features of the initial CV of the target word. The primary consequence of this looser cohort match is an increase in competitor set size: clearly more lexical items will match to a $C_{loose}V$ or CV_{loose} structure than to the strict, more constraining, CV description.

This looser matching constraint has consequences for the potential effects of phonological reduction processes on competition levels. Consider what happens when the final segment in *line* is assimilated. Assimilation to a following labial results in lexical ambiguity: [laɪm] matches to *lime* as well as to the assimilated *line*. There is no corresponding **strict** CVC competitor to the velar assimilation [laɪŋ]. However, the competitor *lying*, a bisyllabic word in citation form may, in running speech, be articulated with a very weak schwa and sound not unlike the assimilated *line*. Thus it may be that the competitor effects of phonological reduction are more significant for word recognition than traditional theories might suggest.

10.3.1 Effects at word endings

Two of the three processes on which this thesis focuses occur at word boundaries. Independent of whether or not these processes occur post Uniqueness Point (however one wishes to define such a position) it is clearly a truism that word boundary reduction processes occur at a position of relatively high predictability compared with word onsets, especially with respect to longer, polysyllabic words. For example, statistics demonstrate that the 7th or 8th segment of any word in English is likely to be /t/ or /d/, independent of the nature of the segments that precede it (Shillcock, personal communication). Does this mean that deleting word-final /d/ or assimilating word-final nasals has little effect on levels of lexical competition?

10.3.1.1 Word-final /d/-deletion

The consequences of word-final /d/-deletion on competition is dependent on at least two factors:

- the length of the word (in terms of segments);
- the morphological character of the /d/ segment.

As observed above, the longer the word, the more predictable the word-final segment identity. Thus the /d/ at the end of *diamond* or *submerged* is more redundant than the /d/ at the end of *gold* or *manned.* In a traditional cohort analysis, deleting /d/ at the end of *gold* will lead to the acoustic input matching better to the competitor *goal* than to the target *gold* (see Example (10.2)). Similarly deleting the final segment of *manned* results in a better match to *man* (see Example (10.3)).

(10.2) [goʊld] – [d] ⟶ [goʊl]

(10.3) [mand] – [d] ⟶ [man]

The pattern of RWRs to the stimulus *gold* reflects this competitor effect, with over one third of RWRs (35/90) being either *go* or *goal.* If a division is made between those tokens with a /d/-duration of less than 60 ms (N=8), and those where the duration of the stop is greater than 60 ms (N=7), it becomes apparent that the /d/-less RWRs (like *go* and *goal*) occur almost exclusively for the tokens with short stop durations (there is just one *goal* response to a long stop token). Conversely, almost all responses to the tokens with longer stops are either perfectly correct (i.e. *gold*) or are words that end in /d/, such as *gored, board* and *rolled.*

When a looser match to onset is accepted, then, of course, the level of competition increases. For example, we find responses of *brown* to the target *round*, and of *car* and *curve* to the target *carved*, all of which can be accounted for by way of $C_{loose}V$ and CV_{loose} matches respectively. Similarly, we find responses of *issue, tissue, tissues* and *dishes* to the target stimulus *disused.* Again, when the tokens are grouped according to the duration of the final stop, virtually all the non-/d/ competitors are found to be responses to the tokens with stops of less than 55 ms duration. In like manner the *trebled* responses to the target *pebbled* occur only in targets where the stop duration is greater than 60 ms.

The stimuli with word-final deletable /d/ fall into two categories: those where the /d/ is an integral part of the stem, such as *diamond* and *round*, and those where the /d/ forms part of the past tense -ED affix, like *carved* and *cobbled*. In natural connected speech, the past tense affix will frequently be predictable from information about tense provided earlier. There is evidence to suggest that, in context, recognising the stem should be sufficient to activate the inflected form (Tyler *et al.*, 1993). In such cases, the deletion of the word-final /d/ should be less detrimental to the recognition process than when the /d/ forms part of the stem itself: it may be that the acoustic information relating to the /d/ is simply not required, the morphological information which is supplied top-down being sufficient for the input to be recognised as, for example, *carved* rather than *carve*.

Recall that one of the problems encountered in defining Uniqueness Points was the question of what to do with inflected forms. Should one define lexical competitors in terms only of base-forms or should all inflected forms be included? Clearly the answer chosen will have significant consequences on the overall level of competition. In word identification tasks such as the intelligibility experiments discussed here, there is no contextual information available to the listener (intelligibility subject). With no reason to assume that all stimuli will be exclusively stem-words only, listeners will known that they're hearing *disused* as opposed to *disuse* only when and if they perceive the final stop.

In terms of the sorts of analyses being undertaken here, therefore, there should be no difference in responses for the two word-final stop conditions, except as the responses relate to the reduction in stop duration. Were /d/-affixed words found to be more frequently hypo-articulated, then one would expect to find more non-/d/ responses to this set of words. It transpires, however, that there is no clear difference in likelihood of hypo-articulating: many of the shortest stops occur in main-stem /d/ words such as *diamond* and *gold*, while some of the longest stops occur in the /d/-affixed words *carved* and *owned*.

This finding offers little comfort to Lindblom's theory of Hypo- and Hyper-articulation. According to the H & H view, /d/-affixed words should have more predictable /d/ segments: the appropriate information regarding TENSE will be supplied by syntactic context, so once the stem has been accessed the word should be recognised and the stop segment 'supplied' top-down, whether or not the acoustic correlates of the stop are actually perceived. Given this redundancy, /d/-affixed words ought to show evidence of increased hypo-articulation, compared with the articulation of word-final /d/ segments that form part of the word

stem. They do not. Speakers are as likely to hypo-articulate words like *round* and *gold* as they are *poisoned* and *reclaimed.*

10.3.1.2 Place assimilation of word-final nasals

The process of assimilation has potentially rather different consequences for lexical competition from those of deletion. It was observed above that deleting word-final alveolars will often increase competition simply as a result of there being less phonological material available for matching: fewer segments means less opportunity to rule out competitors that mismatch. Assimilation, on the other hand, retains the number of segments, altering, instead, the featural description of the segment undergoing the process of change. Thus the effects on competition relate to the availability in the lexicon of similar words to the target, but which end in /m/ or /ŋ/ rather than /n/, for example.

Is there any evidence from the RWRs that assimilating the word-final nasal alters the competitor set for the target? A strict left-to-right segment matching procedure like that advocated by early versions of the Cohort model would predict that word-final nasal assimilation has no effect on lexical competition for the set of target words used in this study. There are no words in English that share exactly the same segmental structure as the targets except for the place feature of the final nasal[8]. For example, *crane* has no competitor with the segmental structure /kreɪm/ (e.g. *crame* or *craim*), while *pine* has no /paɪŋ/ competitor.

However, the pattern of RWRs reflects a less strict matching of featural information. In the sense that RWRs which fail to match the first CV are not correct recognitions of the target word, the Cohort model is, of course, accurate. But in the sense that a stimulus word – that is, the acoustic input – does not always call forth a perfect match, the model is wrong. It fails to capture the location and, possibly, cause, of featural mismatch. So, for example, when subjects respond to the target *crane* with a competitor like *cream*, or to *pine* with *paying*, the strict cohort approach misses the connection between the input stimulus and the Real Word Response.

A loose cohort matching procedure, on the other hand, which permits less exact matches for place information, can capture the relation between /kreɪn/ and /krim/ by a match to [krV$_{tense}$C$_{nasal}$][9]. Thus the responses of *cream* to *crane* can

[8]Clearly this is not true of the lexicon in general, as evidenced by minimal pairs such as *feign/fame* and *kin/king.*

[9]In fact, the broad category V$_{tense}$ itself fails to capture the similarity between the two

be seen to be more predictable than their failure to match at CV onset would at first suggest.

As in the /d/-deletion analysis, a significant factor appears to be the role of morphological inflections. If the level of lexical competition is assessed purely on word-stems alone, then assimilating from alveolar to either labial or velar involves a change to a much smaller word-final nasal dataset: while there are 2337 words in the CELEX LEMMA database that end in /n/, there are only 1156 /m/-final words and 1227 /ŋ/-final words. If, however, competition is based on an analysis of all word forms, i.e. including inflections, there is a significant increase in /ŋ/-final words to compete with the assimilated target, because of the productive -ING affix in English. Thus, while the number of /n/-final and /m/-final words remain relatively stable (2638 and 1169 respectively) the number of /ŋ/-final words rises to 6496 (see Table 10.20).

Word-final match	**CELEX database**	
	LEMMA	WORD
/n/	2337	2638
/m/	1156	1167
/ŋ/	1227	6496

Table 10.20: Number of words in CELEX database that end in different nasal segments

In a preliminary investigation of intelligibility responses to nasal words described in Chapter 7 it was observed that although there was a correlation between phoneticians' judgements of [m]-ness and RWRs ending in /m/, there was not an identical relation between judgements of [ŋ]-ness and RWRs ending in /ŋ/. Judgements of [ŋ]-ness correlated with /ŋ/-final responses only preceding velars, but not labials. It was hypothesised that subjects might be responding with /ŋ/-final words regardless of the quality of the nasal segment, because the productive -ING affix simply generates so many /ŋ/-final competitors. However subject responses rarely ended in /ŋ/ when there was no perceptual evidence of assimilation. It would seem that although there are many /ŋ/-final competitors available, they require an appropriately assimilated production before they successfully challenge the target word for recognition.

What the distribution of nasal-final words in the lexicon tells us is that when such

vowels for both RP and Scots: while in RP the vowel /i/ is very close to the end-point of the diphthong /eɪ/, in Scots the vowels /i/ and /e/ are adjacent in the vowel space, and differ only in the degree of closeness.

a non-canonical production is encountered – i.e. when speakers hypo-articulate the word-final nasal segment – the size of the competitor set for labial and velar nasals is significantly different ($N = 15021, \chi^2 = 1790, p < .001$). There are five times as many words ending in /ŋ/ as there are words ending in /m/. In other words, there are more /ŋ/-final words available in the lexicon to compete with a pre-velar assimilated lexical /n/, than there are /m/-final words, reducing the chances of recognising a pre-velar target correctly. Indeed, 12 of the 19 nasal target stimuli are found to have more /ŋ/-ending competitors in their loose cohort than /n/-ending words; there are never more /m/-ending cohort members than /n/-ending words.

Some of the target words are found to have close competitors in the loose cohort which are offered most often as an incorrect Real Word Response. For example, subjects offer *lemming* for *lemon*, *ending* for *Indian*, and *overgrowing* for *overgrown*. In no cases are there dominant RWRs that end in /m/ for nasals preceding labials (e.g. *Crane*, *green Roman*, *seven*). Responses to the target *fallen* which occurred both pre-labial and pre-velar illustrate the different competitor effects of place. When *fallen* precedes *pillars* the effect of assimilation is negligible. There are no /m/-final competitors, assimilation scores are high, but so, too are the recognition rates. When *fallen* precedes *cairn*, on the other hand, despite much lower judgements of assimilation (mean assimilation pre-velar = 0.417; pre-labial = 2.111) there are many more recognition errors, with *falling* dominating the response profile.

It would seem, then, that speakers can afford to assimilate pre-labial nasals because the effect on competitor size will be small. Pre-velar assimilation does not afford speakers the same luxury because of the competition presented primarily by -ING-final responses.

10.3.2 Effects at word beginnings: schwa syncope

One implication which arises from the finding that subjects respond to polysyllabic stimuli with word candidates of a different – frequently shorter – syllabic structure (see Section 10.2.2 above), is that metrically Weak syllables fail to be perceived accurately. Of course given a principle of articulatory economy (Lindblom, 1990a), and the observation that the unstressed vowels in Weak syllables are frequently deleted (Dalby, 1984), it is not clear whether the failure to recognise Weak vowels arises from a difficulty in hearing acoustically weak cues, or a problem of identifying what simply is not in the signal to begin with.

The possible deletion of Weak vowels clearly has implications for the process of lexical access. As Dalby (1984) observes, schwa syncope usually requires a process of resyllabification. Post-stress syncope frequently results in sonorant consonants adopting the role of syllable nucleus, as when *button* is pronounced [bʌtn̩]. The deletion of pre-stress schwa in words like *banana* and *saloon*, on the other hand, tends to give rise to consonant clusters of varying degrees of 'acceptability' or 'pronounceability', such as [bn] versus [sl].

What, then, is the likelihood of schwa being deleted in pre-stress position? Dalby argues that the incidence of schwa syncope correlates positively with the pronounceability of the newly formed cluster – predicting that schwa is more likely to be deleted in productions of *saloon* than *banana* – but these are exactly the cases that are dangerous in terms of word recognition, since the new cluster will activate an entirely different cohort of lexical competitors. Deleting the schwa in *saloon*, for example, would lead to the activation of /sl/-initial words such as *sluice, slew* and *sloop*, as opposed to *saloon* and *salute.*

Indeed, the strong version of the Cohort model would simply fail to activate the Weak-initial candidates. It is not clear that more recent instantiations of the Cohort model (Lahiri and Marslen-Wilson, 1991; Gaskell and Marslen-Wilson, 1996) can cope any better with the full deletion of Weak onsets except via a process of phonological inference if and when mismatch is encountered. The word-initial cohort activated for the reduced token [slun] cannot offer a lexical match to the nasal input: there is no word in English for which the stored citation-based representation starts /#slun/ (or indeed /#slum/ or /#sluŋ/). The nasalisation, in this instance, might serve to prompt listeners to employ a process of phonological inference and posit an 'underlying' Weak vowel between the /s/ and /l/, which the speaker has deleted. In this way, listeners might activate a revised cohort of /səlu/ words, and proceed to recognise the input.

A model like TRACE, on the other hand, can account for the recognition of *saloon* from [slun] on the basis of best overall match. While Weak-initial words like *saloon* will receive early activation from the initial segment /s/, the deletion of schwa will result in competitors like *sluice* and *sloop* inhibiting the target word, dampening down the activation levels of *saloon* and *salute.* However, once the feature [+nasal] is activated, the spread of activation will boost the level of *saloon* and eventually *saloon* will emerge as the best fitting candidate.

TARGET	DUR(ms)	RESPONSES
saloon	368.37	- 1 chillen 1 1 - 1 1 1
saloon	305.62	slim cloud - slim slim time tellyroom plough -
saloon	488.13	1 selling 1 1 silly 1 sailing 1 -
saloon	413.69	swim swim swan sun salon swim swim swim swan
saloon	388.44	swimming aloon 1 1 - 1 - salute 1
saloon	681.56	1 1 selnium - 1 - 1 1 1
saloon	407.31	slow sewing salon phone sowing swimming 1* 1 silent
saloon	368.25	numb ttollm - slim tulip - swim ceiling chillum
saloon	560.81	sullen filling syringe syringe 1 wind 1 swimming syringe
saloon	530.31	1+ 1 1 1 1 1 1 1 1
collapsed	326.37	1. 1. 1 clutch clapped twat class clerks claps
collapsed	307.44	twerp clap - claps clap clap clasp 1 claps
collapsed	450.94	1. 1 1 1 1. 1 1. 1. 1.

Table 10.21: Responses to the stimulus words *saloon* and *collapsed*, showing evidence of possible resyllabification after pre-stress schwa syncope

10.3.2.1 Listener responses to WS stimuli which can resyllabify

Is there any evidence that subject responses are directed towards the word-initial cohort associated with the cluster than derives from the deletion of schwa in WS polysyllables? Unfortunately, only two WS polysyllables which have acceptable clusters when resyllabified were used in the intelligibility studies: *saloon* and *collapsed*. The full set of responses to these two words can be found in Table 10.21.

As can be seen, RWRs to *saloon* included *slim*(4) and *slow*(1), while RWRs to *collapsed* included *clap*(3), *claps*(3), *clapped, clutch, class, clasp* and *clerks*. Indeed, of the 13 RWRs to the stimulus *collapsed*, only two did *not* start with a /kl/ cluster, and of the /kl/ responses themselves all but one (*clutch*) had an appropriate [a]/[ɑ] vowel. In addition to the /#sl/ responses to *saloon*, subjects also matched the input to *swim*(6), *swimming*(3) and *swan*(2). These /#sw/ responses are predictable, given the broad class categories of the loose cohort matches.

It is plausible, then, that reduction of pre-stress schwa in polysyllabic WS words which could resyllabify into acceptable clusters might affect subjects' ability to recognise the word. Clearly the data presented here is insufficient to warrant 'evidence' but it entertains the possibility that subjects may indeed be directed towards competitors that share the same onset as the cluster that would derive from the deletion of schwa in certain WS polysyllables.

10.3.2.2 Assessing the potential for lexical competition after resyllabification of WS onsets

To ascertain how frequently pre-stress schwa syncope might result in lexically ambiguous onsets, the CELEX database was searched for all words with a Weak-initial syllable. The search was restricted to all words starting (C)(C)ə; Weak syllables containing /ɪ/ were excluded on the grounds that it is not clear whether /ɪ/ is ever fully deleted (as opposed to being *reduced* in duration and quality only). The 1737 words returned by this search vary in the acceptability of their onset when the schwa is fully deleted, with only 360 WS polysyllables having onsets which form phonotactically permissible cluster sequences after schwa deletion. The set of word-initial clusters which match at least one word in the CELEX database is presented in Table 10.22 along with an indication of how many competitors start with the same onset.

Of the 34 clusters, 10 begin fewer than four lexical items. These can be discounted on the grounds that they are either words clearly loaned from another language and pronounced according to the phonology of the source language, such as *svelte* (pronounced /svɛlt/) or *schnapps* (pronounced /ʃnaps/), or because they are errors in transcription, such as *pseudo* (transcribed /psjudoʊ/), and *pooh* (transcribed /phu/). The remaining 24 clusters between them elicit a total of 6064 competitors, with the minimum cohort size being 7 (for both /#sf/ and /#skj/) and the maximum 880 (for /#pr/).

This is the number of competitors at the consonantal onset. A traditional CV cohort approach would assume, of course, that the following stressed vowel is correctly identified; in this case the number of competitors falls considerably. For example, although there are 246 words which start with the cluster onset /#bl/ (derived from deletion of schwa in *balloon*), only 24 of these words share the vowel /u/: *blue, bloom etc.* By way of illustration, Table 10.23 shows the potential number of competitor words when schwa is deleted from the onset of WS stimuli that start with /#səl/, such as *saloon.*

Such an analysis gives the impression that schwa syncope may not frequently result in the sort of lexical ambiguities with which we are currently concerned. But we know from the discussion in Section 10.2.3 that the expectation that subjects successfully match the acoustic input to at least the first CV of the target is over optimistic. If a looser-based matching criterion is advocated, then clearly the level of competition will rise. In the example presented in Table 10.23 above, for instance, a match to the looser CV_{loose} structure would result in 111

Onset cluster	Cohort size	Obscure/exceptional cases			
S.m.	2	schmaltz	S.m.Q.l.t.s.	schmaltzy	S.m.O:.l.t.s.I.
S.n.	2	schnapps	S.n.{.p.s.	schnitzel	S.n.I.t.s.l,.
S.r.	27				
b.l.	246				
b.r.	334				
f.l.	281				
f.r.	285				
g.l.	120				
g.r.	316				
k.l.	288				
k.m.	1	Khmer	k.m.E@.r*.		
k.n.	1	Knesset	k.n.E.s.@.t.		
k.r.	398				
k.v.	1	kvass	k.v.A:.s.		
m.j.	58				
p.f.	1	pfennig	p.f.E.n.I.g.		
p.h.	1	pooh	p.h.u:.		
p.l.	253				
p.r.	880				
p.s.	3	pseudo	p.s.j.u:.d.@U.	psoriasis	p.s.Q.r.aI.@.s.I.s.
		(3rd case is *psst* transcribed as p.s.)			
s.f.	7				
s.k.	444				
s.l.	219				
s.m.	93				
s.n.	128				
s.p.	399				
s.r.	1	Sri	s.r.i:.		
s.t.	716				
s.v.	1	svelte	s.v.E.l.t.		
t.r.	473				
t.w.	55				
s.k.j.	7				
s.p.r.	48				
s.p.l.	30				

Table 10.22: Possible onset clusters in English that arise from the deletion of schwa in pre-stress position, with size of lexical competitor set, and listed obscurities

Onset of WS word	New cluster	[s.l.]V words (no.)	WS words (no.)	Transcription of WS word	Orthog	Freq (Cob 17.9m)
				(Examples of new cohort words)		
s.@.l.A:.	s.l.A:.	11	2	s.@.l.A:.m.I.	salami	25
				(*slant, Slav, slanderous, slalom*)		
s.@.l.E.	s.l.E.	7	1	s.@.l.E.m.n.@.t.I.	solemnity	35
				(*slender, sledge, sled, sledgehammer*)		
s.@.l.I.	s.l.I.	38	14	s.@.l.I.s.I.t.@.r*.	solicitor	387
				(*slip, slim, sling, slipper, slither*)		
s.@.l.aI.	s.l.aI.	14	1	s.@.l.aI.v.@.	saliva	52
				(*slightly, slide, slice, sly, slimy*)		
s.@.l.eI.	s.l.eI.	18	3	s.@.l.eI.S.@.s.	salacious	4
				(*slave, slavery, slate, sleigh, slake*)		
s.@.l.u:.	s.l.u:.	9	6	s.@.l.u:.S.n,.	solution	1332
				(*sluice, slew, sloop, sleuth*)		
s.@.l.{.	s.l.{.	27	1	s.@.l.{.s.@.t.I.	salacity	2
				(*slap, slam, slash, slacken, slang*)		

Table 10.23: Lexical competitors to [#səl] initial words when schwa is deleted

competitors with [#slV$_{lax}$] onsets and 108 competitors with [#slV$_{tense}$] onsets.

Recall how in Section 10.3.2.1 subjects were shown to respond to the target stimulus *saloon* with RWRs such as *swim* and *swan*. If the matching criterion for *saloon* is loosened to include any [#sC$_{[rlwj]}$V] onset then the competitor set rises to 366, of which 78 have the form [#sC$_{[rlwj]}$VC$_{[nasal]}$].

Schwa syncope, then, may be a greater problem for lexical competition than might appear at first glance. Hypo-articulating pre-stress schwa may prove lexically dangerous because it leads listeners to access inappropriate cohorts of competitors. Because the phonological reduction occurs early in the word, the predictability of the segment to follow – or redundancy – is low; the application of the process may therefore be more problematic than would be, for example, the deletion of word-final /d/.

10.4 Lexical competition and the H & H theory

Recall that the H & H theory predicts that speakers reduce articulatory effort according to the informational needs of their listener: where tokens are predictable from context, for example, speakers can afford to hypo-articulate. The constraint imposed on speakers to prevent them from hypo-articulating to extremes of un-

intelligibility is one of 'distinctiveness': speakers can hypo-articulate only while lexical distinctiveness is maintained. I interpret this to mean that speakers can hypo-articulate so long as listeners are able to distinguish the target word from other lexical competitors. In the early sections of this chapter I discussed how one might define 'lexical competitor' when dealing with tokens of natural unscripted connected speech. In addition, I observed what effects connected speech processes might have on the size of the lexical competitor set: for some words altering the word-final segment by deletion or assimilation introduces lexical ambiguity, while for other words there are no lexical consequences associated with such changes. Similarly, the deletion of pre-stress schwa can have potentially dangerous consequences for the listener, in terms of his ability to decode the input.

A theory that advocates a consideration of listener needs, and which, furthermore, relies on a principle of lexical distinctiveness to curb articulatory economy (Lindblom, 1990a), might therefore predict – contrary to Dalby (1984), for example – that speakers will refrain from deleting pre-stress schwa or assimilating word-final nasals in exactly these high competitor situations, that is, where syncope or assimilation leads to possible lexical ambiguity. In other words, the H & H theory would predict:

Hypothesis 10.2 *Where nasal assimilation has no effect on the size of the lexical competitor set, speakers may assimilate without affecting recognition responses. However, speakers should refrain from hypo-articulating introductory mentions where assimilation results in the activation of close (and therefore dangerous) lexical competitors.*

Hypothesis 10.3 *Where schwa syncope results in a phonotactically impermissible cluster, it will have no significant effect on recognition responses; in such cases speakers can delete schwa without increasing the lexical competitor set. Conversely, where deletion of schwa leads to phonotactically acceptable onsets, speakers should refrain from hypo-articulating to avoid activation of inappropriate competitors, at least when introducing the word into the discourse for the first time.*

In other words, if speakers are sensitive to their listeners' informational requirements, then they should refrain from articulatory reduction when lexical competition is high, but produce 'poor' tokens when there are few competitors around to make recognition difficult.

To test this hypothesis, the set of nasal-final target stimuli were classified according to whether or not they had close matching competitors that arise from the application of the assimilation process. Loose competitor sets were generated for each stimulus word, using regular expression searches on the CELEX database, based on the broad class feature matches discussed earlier in Section 10.2.3 of this chapter. Each competitor set was then searched for the number of competitors that ended in the appropriate nasal when the word-final /n/ of the target stimulus was assimilated. Hence, the loose cohort competitor set for words like *crane* and *saloon* were searched for /m/-ending words, while the cohorts for *fallen* and *Indian* were checked for /ŋ/-final words. The target stimuli were then assigned to one of two groups: +/− large competitor set, according to the size of the competitor set, or number of **Close Competitors** (CCs). The proportion of nasal-final competitors to the competitor set as a whole was computed, the proportions ranked according to size, and the stimulus set then divided in the middle by the median.

As a check on the appropriateness of dividing the stimulus set in this way, a separate, independent division of the data was made on the basis of the dominant RWRs to the stimuli in the intelligibility experiments. For example, the prevalence of the RWR *ending* to the target stimulus *Indian*, of *falling* to *fallen*, and of *swim* to *swan* meant these words clearly had close competitors to contend with, while the absence of any dominant nasal-final response to targets such as *seven* and *pine* classed these as targets with no competitors to fight off. The results of the two analyses showed the same effects.

The predicted outcome based on Hypothesis 10.2 can be seen in Table 10.24. The number of close competitors should have no or little effect on the likelihood of assimilation for second mentions, since reference to a previously mentioned entity could allow listeners to recognise the input with the help of their mental representation of the discourse. The word's Givenness should allow speakers to hypo-articulate without a detrimental effect on their listeners' ability to recognise the word. However, there is no such contextual aid to recognition for introductory mentions. Therefore, where the number of close competitors is high, speakers are predicted to refrain from hypo-articulating, since a poor first mention of a word with many similar sounding competitors will be potentially dangerous for their listener.

An ANOVA was run with degree of assimilation as the dependent variable, and mention (First/Second) as a repeated measure. The size of the competitor set

(presence/absence of lexical competitors) was a between-items grouping factor. The means for each condition can be found in Table 10.25.

	Competitor set	
Mention	+CCs	−CCs
First	dangerous	OK
Second	OK	OK

Table 10.24: Relative danger of assimilation for first and second mentions according to the presence or absence of Close Competitors (CCs) once assimilated

	Competitor set	
Mention	+CCs	−CCs
First	1.840	1.500
Second	1.840	1.438
N	*16*	*18*

Table 10.25: Perceived assimilation for first and second mentions according to the presence or absence of Close Competitors (CCs) once assimilated

No effect of competitor size was found ($F_1(1,16) = 1.86, p = .19$; $F_2(1,32) = 1.06, p = .3$), nor was there any significant effect of mention ($F_1 < 1, F_2 < 1, n.s.$). Speakers assimilated first and second tokens equally, with no allowance made for the potential danger of close competitors even when introducing landmarks into the discourse for the first time. Indeed, the means show a slight but non-significant increase in judged assimilation for those words with close competitors, rather than a decrease.

Contrary to the predictions of the H & H theory, then, speakers appear to hypo-articulate – illustrated here by the assimilation of word-final nasals – irrespective of whether, in doing so, they make their listener's task of recognition a more difficult one. This finding appears to be counter to Lindblom's claim that speakers adjust variability in production according to an assessment of their listeners' needs. It would seem, rather, that speakers pay scant regard of the effects of hypo-articulation on their listeners' ability to recognise what they're hearing.

In Lindblom's defence it could be argued that an analysis of words excerpted from context and presented to subjects to recognise without the additional material available to the original listener fails to acknowledge the contribution made by "signal-independent sources". Isolated word recognition tasks, such as the intelligibility experiments presented in this thesis, enable us to explore what acoustic information is or is not available in the signal, independent of how predictable

the word might be. (Cloze tests on the utterances from which they were taken would provide a measure of the token's predictability were it required.) The experiments provide a measure of how easy a particular token is to recognise, and what other lexical items listeners think the token sounds like. What is at issue, here, is the processing consequences of reduction at a *lexical* level. The H & H theory argues for a distinctiveness constraint that 'prevents' speakers from hypo-articulating when *words* are no longer distinctive. What makes a word distinctive is a function of what other words are available in the lexicon that share features with the target, in other words, its competitors. The evidence presented in this chapter suggests that lexical competition does not appear to influence the degree to which speakers hypo-articulate introductory mentions: they assimilate even when doing so will result in the activation of a competing word. Clearly there must be a constraint on hypo-articulation at some level – without one the content of a speaker's message would not be recognised – it just may not be a *lexical* constraint.

10.5 Discussion

I have suggested that the pattern of incorrect responses given in the series of intelligibility experiments provides indirect but potentially useful information about lexical processing. Instead of assuming that the activation of multiple lexical candidates is based on the correct matching of featural information in temporal sequence, establishing word-initial cohorts based on CV onset for example (Marslen-Wilson and Tyler, 1980), the lexical competitor set can be operationally defined by means of the set of alternative candidates offered by subjects in response to the stimulus.

The degree and location of feature match between the stimulus input and the subject responses is interpreted as reflecting the **relative salience** of different parts, or segments, of the word. Because onsets are **less redundant** than word offsets it is likely that in the majority of cases speakers will produce relatively clear word beginnings, increasing hypo-articulation through the word. But since some parts of the acoustic stream are more perceptually dominant than others, the location of match and mismatch will be a function not just of redundancy, but of **acoustic identity**.

Thus strict CV onset matches are rejected in favour of matches to a cohort of competitors which share some rather than all of the featural information in the

onset. I argue that target stimuli match to CV_{loose} or $C_{loose}V$ competitor cohorts – where 'loose' refers to a broad manner-based class of segments – according to the phonetic characteristics of the word onset: CV_{loose} cohorts are activated by stimulus words starting with voiceless stops or fricatives and words with lax vowels; $C_{loose}V$ cohorts are activated by initial voiced stops, words with consonant cluster onsets, and words with tense vowels. In addition, a distinctive affricate (or [ks] cluster) might dominate the pattern of responses to target words like *village* (or *Saxon*) independent of other factors, because of the strong acoustic cues to segment identity at this point.

As was observed in Section 10.3 a likely (and probably inevitable) consequence of this looser cohort match is not simply the introduction of a **different** set of lexical competitors but an **increase in competitor set size**: in general, more lexical items will match to a CV_{loose} or $C_{loose}V$ structure than to the more constraining CV description. This prediction would hold true for both traditional Cohort-based and TRACE-like models of word recognition. In the former case the initial contact cohort is larger; in the latter, looser matching will reduce the amount of lateral inhibition since there will be less acoustic information feeding in at the featural level to inhibit the large set of activated candidates.

Any predicted expansion in competitor set size increases the likelihood of phonological reduction processes having a significant effect on lexical processing. But it also raises the question of just how large a competitor set is activated: does the looser matching criterion introduce too large and unwieldy a set of competitors to be useful, or even plausible, as a lexical processing strategy?

There are several observations to be made here.

The first concerns the nature of the input representation and the kind of information likely to be relevant for successful spoken word recognition. I see the Loose Cohort model as an attempt to move away from lexical processing models that require a level of *phonemic* representation towards a model of spoken word recognition that responds to perceptually salient phonetic features. This idea is not new; Klatt (1989) argued against any intermediate level of representation mediating between the acoustic input and lexical representation. Indeed, one of the most recent versions of the Cohort model (Gaskell and Marslen-Wilson, 1995) advocates direct access from featural information to lexical representation. But the Loose Cohort model differs in that it implies a possible **weighting** of featural information, according to its acoustic/perceptual salience, allowing for looser matches to less distinct parts of the input. It should perhaps be noted here

that although the TRACE architecture utilises a phonemic level of encoding the fact and nature of this intermediate level is not critical to the functioning of the system: a TRACE-like model could in principle incorporate a more direct relation between featural description and lexical representation. Clearly it is plausible and even likely (though not essential) that the underlying lexical representation is similarly couched in featural rather than phonemic terms.

Secondly, there is no *a priori* requirement for the weighting associated with perceptual salience to be fixed or permanent; thus it is possible to envisage a model of lexical access where the looseness of match between input and response varies dynamically in relation to how carefully articulated the production is: hypo-articulated forms will be associated with a greater degree of match looseness, hyper-articulated forms with a stricter match. Such a model would generate competitor sets which varied in size according to the acoustic information made available by the speaker; in other words, the competitor set would expand and contract in proportion to the token's clarity. In Section 3.2.2 it was observed that the richer and more discriminative the information in the input representation, the smaller the number of lexical entries initially contacted. The Loose Cohort model proposes that the degree to which the input representation abstracts away from the acoustic detail will vary, according to the level of articulatory clarity, rather than being fixed.

It may be recalled (see Section 6.7) that there is a relation between linguistic form and the cognitive status of discourse entities: the degree of attenuation in linguistic production functions as a marker of accessibility, with reduced tokens signalling to the listener that the referent ought to be amongst the most salient in his discourse representation. A large set of loosely matched competitors ought similarly to signal to the listener that he should search the set of previously mentioned entities for a suitable candidate. The discourse representation then functions as a **filter** through which the set of lexical competitors must pass. In other words, although the Loose Cohort model might generate a larger competitor set than traditional models, the increase in size might itself be informative, aiding rather than hindering the recognition process.

A related issue is that of the role of context in determining the lexical competitor set. It is not yet clear what influence – if any – context may have in determining which lexical competitors are activated. The results from various gating experiments are contradictory: while Tyler and Wessels (1983) find that listeners' guesses at early gates are often contextually inappropriate, McAllister

(1988) showed that context *does* have an effect on initial guesses. If further research proves McAllister (1988) to be right, then the larger cohorts activated by the Loose Cohort model might be reduced by contextual appropriateness, thus removing the problem of unwieldiness.

Similarly, the results of priming experiments on assimilated tokens uttered in context (Gaskell and Marslen-Wilson, 1996) suggest that the difficulties associated with recognising words in isolation are not so relevant when dealing with the kind of speech data we naturally encounter. Within the context of a situated discourse it is possible that the large cohorts generated by a Loose Cohort model would be too short-lived for their size to be problematic.

Finally, there is evidence to suggest that any disadvantage associated with a large and unwieldy set of lexical competitors may be mitigated by the Markedness Ordering Principle (Shillcock *et al.*, 1996). Shillcock *et al.* observe that there is a clear tendency for phonological change like assimilation to introduce a more marked (less frequently occurring) segment than the earlier (pre-change) form. Thus the more frequently occurring /n/ assimilates to the less frequent /m/ or /ŋ/. The Markedness Ordering Principle has been shown to have the functional advantage of flattening the information redundancy curve. Given this observation, it is likely that the increase in competitor set size associated with loose matching to phonologically reduced tokens will be relatively small, compared with the set of competitors generated by a stricter match to the non-reduced form.

The Loose Cohort model, then, is a dynamic model involving a constantly varying set of lexical competitors. A question central to this thesis is how the expanding/contracting competitor set relates to Lindblom's H & H theory of articulatory variability. In theory, the set of lexical competitors will expand as speakers hypo-articulate and contract as speakers hyper-articulate. Specifically, the H & H theory predicts that speakers hypo-articulate **so long as listeners can still distinguish the word target from other competing lexical items**. Where the level of lexical competition is high, therefore, speakers ought to refrain from hypo-articulated productions, at least for introductory mentions when there is no previously established discourse representation to aid recognition.

An analysis of perceived assimilation for first and second mentions of nasal-final target stimuli which did or did not have close competing competitors ending in the assimilated nasal demonstrated that speakers assimilate introductory mentions no less than they assimilate repeated mentions, and that the presence/absence of potential competitors has no effect on the level of assimilation for either mention.

The degree of hypo-articulation, then, is independent of any kind of assessment of potential lexical competition. In other words, speakers hypo-articulate even when doing so results in a production that is lexically ambiguous; that is, speakers fail to maintain distinctiveness.

In the light of this finding, it seems likely that the successful recognition of hypo-articulated forms will depend on listeners actively employing higher-level discourse information to predict the most plausible candidate, given what they already know. While such a view is far from counter to Lindblom's general intuitions about the process of recognising speech, it clearly contradicts the prediction made by the H & H theory that the constraint on speakers' level of hypo-articulation is a **lexical** one.

Chapter 11

Summary and Conclusions

11.1 Introduction

In this chapter I summarise the research findings of the previous chapters and discuss their implications for Lindblom's theory of Hyper- and Hypo-articulation. Finally I highlight some potential areas for future research which could be undertaken to extend this work further.

11.2 Summary of research findings

At the outset of this thesis I presented an account of Lindblom's theory of Hyper- and Hypo-articulation – the H & H theory – which Lindblom offers as an explanation for the variation in production that characterises natural connected speech. The H & H theory argues for a speaker who specifies phonetic structure in the acoustic signal to the extent that it is needed by her listener to supplement signal-independent contextual knowledge: it is a theory about speaker choice in relation to listener need. Under this view speech motor control is teleologically organised, or purpose driven, with articulatory production being tailored to the communicative goals of the discourse.

Central to the H & H position is the principle of economy of effort: speakers will naturally gravitate towards simplification and reduction, as a consequence of rate and synergy constraints on speech production. Lindblom argues that consonant-vowel coarticulation and vowel reduction both illustrate the economy of effort principle in operation. The teleological component to the H & H theory prevents speakers from economising beyond the point at which their listener can recognise the message; a distinctiveness constraint acts as a check on articulatory

reduction, with lexical contrast the key to whether a speaker can afford to reduce clarity and hypo-articulate. Lindblom likens the process of speech production to a tug-of-war between the internal system-based pull towards economy on the one hand, and the external listener-oriented demands for recognisable output on the other.

In presenting Lindblom's account I highlighted two 'sins of omission', areas of the theory which require more detailed treatment if the theory is to have any kind of predictive power. First, Lindblom offers little discussion on the notion of distinctiveness: of what the perceptual constraints on production might be. Secondly, the H & H theory fails to make explicit the kind of signal-independent information that might be available to the listener, and how such information might be used to help decode the acoustic signal. I addressed these two problems in the ensuing chapters.

The basic concepts and key research findings concerning the cognitive process of spoken word recognition were presented in Chapter 3 as an introduction to the treatment of distinctiveness. What makes one word distinct from another depends upon the nature of the lexical competition: the number of other candidates in the mental lexicon that share characteristics with the target word. I argued that variability in production affects word recognition to the extent that it changes the nature of the lexical competition.

Chapter 4 discussed the work of various researchers who have explored the notion of information as it relates to linguistic structure. It asked what kind of record a speaker might maintain of her listener's access to and requirements for information. The distinction was highlighted between information that is New to the discourse, and information that is Old, or Given. Some of the linguistic means available to convey the distinction were reviewed along with certain implications for information storage and retrieval in a cognitive representation of discourse.

It was hypothesised that if the H & H theory is correct, then reference to Given information ought to undergo greater reduction – to be produced with increased speaker economy – than reference to something New to the discourse. This is because in the former case the listener has access to a representation of the entity already established in his discourse model. A series of intelligibility experiments, described in Chapter 6, explored the effects of available information on subjects' ability to recognise tokens of single words excerpted from unscripted Map Task dialogues (Anderson *et al.*, 1991).

The results showed that once an entity is textually evoked by previous men-

tion (Prince, 1981), subsequent reference to the entity is associated with a loss of intelligibility: second mentions of landmark names are less intelligible than their introductory counterparts. This intelligibility loss is not affected by which speaker introduced the entity: same and other-speaker repetitions result in equal intelligibility loss. Nor is it affected by visual access to the referent: repeaters reduce intelligibility even after receiving feedback that their listener cannot see the referent, and when they cannot see the referent themselves.

The representation of an entity in a cognitive model of the discourse clearly contributes to the signal-complementary sources of information that a listener uses to help decode the signal. Thus far the results support Lindblom's view that a speaker reduces articulatory effort when she believes her listener has access to appropriate contextual information that will aid recognition: previous mention leads to reduced tokens when repeated.

The next three chapters (Chapters 7-9) assessed the suitability of certain phonological reduction processes as candidates for the sources of observed unintelligibility: was intelligibility loss associated with increases in word-final assimilation, word-final stop-deletion or pre-stress schwa syncope? Here the results are somewhat equivocal. Although reduction processes are found to be more prevalent in tokens from spontaneous speech than in matched citation forms, they generally fail to account for effects of repetition. Whilst there is an increase in place assimilation preceding velars for repeated mentions, there is no corresponding increase for assimilation preceding labials, nor for the deletion or shortening of either word-final /d/ segments or pre-stress schwa in WS polysyllables. A significant effect of repetition is found, however, for the duration of stressed vowels: lexically stressed vowels are significantly shorter in second mentions of landmark names than in introductory tokens. Reduction in stressed vowel duration, then, mirrors the loss of intelligibility established for repeated tokens.

It was observed that – unlike variation in the production of stressed vowels – connected speech processes which occur at word boundaries, such as place assimilation and stop deletion, may frequently occur after a word has become lexically unique. Perhaps the failure to find a significant correspondence between intelligibility loss and /d/-deletion, for example, arose from the /d/-deletion occurring beyond the word's Uniqueness Point, and therefore not affecting subjects' recognition responses? In Chapter 10 it was hypothesised that variation in speech production will affect recognition success according to the change in competition level arising from the production difference. Where articulatory reduction

introduces lexical ambiguity, the resulting acoustic output should be harder to recognise than the non-reduced version. When reduction makes no difference to competition (such as changes to a segment post Uniqueness Point) the effect on recognition will be small.

When a traditional definition of lexical competitor (Marslen-Wilson and Tyler, 1980) is used, with strict matching to CV onset, the word-final reduction processes are found not to affect the lexical competition of the targets being investigated. For example, there is no lexical match to the assimilated form of *Indian* preceding *country*: [ɪndjəŋ], nor to a token of *round* that has undergone word-final /d/-deletion: [ɾaʊn]. However, an analysis of the responses offered by subjects in the intelligibility experiments reveals that traditional definitions of lexical competition fail to explain the matches subjects make from acoustic input to lexical representation. Many of the responses fail even to access the correct word-initial cohort. An alternative definition of lexical competitor is offered in terms of a 'looser' match based on shared feature rather than strict phonemic matching. Under this definition of competition, the effects of changes at word boundaries become more apparent. For example, the assimilated token of *Indian* elicits responses of *ending* while [ɾaʊn] elicits responses such as *brown.*

Finally, armed with this new definition of lexical competitor, I asked whether lexical competition predicts the likelihood of targets undergoing reduction. An analysis of assimilation revealed that the presence/absence of lexical competitors has no effect on observed levels of assimilation: speakers assimilate introductory tokens of landmark names whether or not the assimilation leads to an acoustic output that activates similar lexical competitors.

11.3 Implications for Lindblom's H & H theory

The H & H theory argues for a speaker who places in the acoustic domain only information that cannot be provided from other sources: speakers keep a running account of their listeners' informational requirements, and economise articulatory effort when they believe their listeners can supplement the information in the acoustic stream with what they already know about the discourse.

The finding that speakers reduce intelligibility when referring back to previously mentioned entities appears to support Lindblom's position. The strong effects of speech form found for the various reduction processes lend further support: speakers reduce articulatory effort, that is, they shorten segments and assimilate more,

in running speech – where there is contextual information to aid recognition – than in carefully produced list readings. However, the effects of repeated mention on reduction processes provide less comfort. The literature on the Given/New distinction and the way it is signalled linguistically suggests that references to Given entities should be more predictable than introductory mentions, and that this predictability should be reflected in linguistic reduction, both in terms of constituent structure and phonology. Our results show only weak effects of repetition on phonological reduction, with some evidence of an increase in assimilation (pre-velar) for second mentions, but no evidence of durational shortening for word-final /d/ or pre-stress schwa.

When reduction is considered in terms of lexical competition, with Lindblom's distinctiveness constraint interpreted as a restriction on articulatory economy which increases lexical ambiguity, no evidence is found of speakers accommodating to their listeners' potential difficulties in lexical access. Assimilation levels are compared for introductory mentions of landmarks which have close competitors ending in the assimilated token (such as *lemon-lemming*), and those with few or no close competitors (such as *seven*); speakers are shown to assimilate the tokens with close competitors to the same degree as tokens with few competitors. Contrary to the predictions of the H & H theory, speakers hypo-articulate irrespective of whether, in doing so, they make their listeners' task of recognition a more difficult one. The degree of hypo-articulation appears to be independent of any kind of assessment of potential lexical competition; in other words, speakers fail to maintain distinctiveness at a lexical level.

Lindblom might argue that the lexical ambiguity presents no problem to the listener in context. For example, knowledge of the previous discourse and of syntax will set-up appropriate expectations for the listener. Thus although the word *beam* might be activated at some point during the process of recognising the phrase "*The eggs have* [bim] *broken*" the listener will encounter no difficulty in reporting that they heard the word *been.* This may be so. But it does not escape the criticism that lexical distinctiveness cannot provide the bottom line for speech economy. Clearly, if speakers are assimilating tokens of *been* in a way that makes them lexically indistinguishable from tokens of *beam* then the lexicon itself cannot disambiguate the production: listeners must call upon processes of inferencing to interpret the input in terms of the most likely candidate, given what has already preceded. While the need for top-down processing does not contradict Lindblom's theory – rather, it would be seen as further supporting it – the necessary conclusion appears to be that economy is not constrained by

lexical distinctiveness. The question that must then be asked is: "What *does* constrain speaker economy?". Clearly some constraint is required that prevents speakers from reducing articulatory effort to the extreme of unintelligibility. If it is not lexical distinctiveness then what is it? How can one define Lindblom's notion of 'sufficient contrast' in a way that makes his theory useful (predictive)? In particular, there is no obvious answer to the question of how clear an introductory mention needs to be, nor to how unclear a subsequent mention can get away with being. Perhaps Lindblom's theory simply boils down to the very general common-sense observations that:

- there's no point in talking if nobody can understand you, but
- people are basically lazy.

Viewed in this light Lindblom's H & H theory is stripped not just of its distinctiveness constraint, but also if its listener-oriented bias. It implies that speakers reduce articulatory effort for purely egotistical reasons – because it makes their own life easier – and will hypo-articulate until they reach the point at which their listener interjects with a "*Pardon? I didn't quite catch that*" response.

Is there evidence that might support this view of a speaker as a less than perfect 'Girl Scout'? Recall that in Experiment Four (Chapter 6) speakers were shown to reduce the intelligibility of introductory tokens of landmark names when they gave instructions about a map they had seen previously to a new listener. Despite the fact that the entity was New for her listener, the speaker produced a less intelligible token than the one she produced in her first encounter with the map. In Chapter 6 we interpreted this finding as evidence of speakers constructing representations of the discourse based on their own experiences; to the extent that a speaker models her listener's discourse at all, the model is based on what the speaker herself knows rather than what she might have cause to believe her listener knows. In many discourse situations the speaker's oversimplification of the genuine state of affairs will not differ significantly from a more accurate model of what her listener actually knows; on some occasions, however, such as when a new Follower does not share the Giver's previous map experience, the two models may differ quite radically, and the simplification may put the listener at a disadvantage.

Lindblom's claim, therefore, that

> *"the speaker estimates the running contribution that signal-complementary processes will make during the course of an utterance, and dynami-*

cally tunes the production of its elements to the short-term demands for either output-oriented control (hyper-speech) or system-oriented control (hypo-speech)" (Lindblom, 1990a, page 405)

ought perhaps to be treated with a certain scepticism. The degree of hypo-articulation associated with a token's production is possibly more egocentrically speaker-oriented than the H & H theory maintains.

However, independent of the motivation for hypo-articulation, the reduction that characterises repeated mentions may offer a practical advantage to the listener. Recall that various researchers (Fowler and Housum, 1987; Terken and Nooteboom, 1987; Bard *et al.*, 1991) have argued that poorly articulated second tokens function as effective primes (in prime-probe experiments, for example) because they require the listener to access their stored mental representation of the discourse in order to recognise the incoming word. In other words, lexical access is facilitated by reduced articulatory precision. If, as Ariel (1990) argues, attenuation signals accessibility then reducing articulatory effort marks the production out as accessible, and constrains a referent's identity to the set of entities previously stored in the discourse representation, consequently reducing the burden of lexical access. Thus, whether or not speakers hypo-articulate with regard to listener needs, listeners may sometimes be able to utilise the fact that a token is poorly articulated to aid recognition. On those occasions, however, when listeners are presented with hypo-articulated tokens first time around, like the Instruction Followers in the second givings of maps (Experiment Four), or the listeners who fail to deny the first tokens of unshared landmarks (Experiment Two), the reduced clarity in articulation will be disadvantageous.

11.4 Some future work

Several aspects of this work warrant further investigation.

In Chapter 7 an asymmetry was observed between judged assimilation of nasals preceding labial contexts and those preceding velar contexts. Although some evidence is offered in Chapter 10 to suggest that this asymmetry might arise from a difference in the size of the close competitor sets containing velar- and labial-final words, the support is not overwhelming. It would be advisable, therefore, to explore the possibility of identifying an articulatory or, indeed, perceptual basis for the difference. This could be done with the use of EPG data to test for evidence of blended as opposed to overlapped articulations, for example. Assimi-

latory articulations could also be produced with varying amounts of overlap and presented in a perceptual study to test whether the timing of gestural overlap had a significant effect on perceived assimilation. Since formant transitions are characteristically longer for velars than labials (Ladefoged, 1982) it may be that duration of overlap contributes to the observed asymmetry.

It was predicted in Chapter 10 (Section 10.4) that if the lexical distinctiveness constraint is correct then speakers should refrain from articulatory economy where phonological reduction would increase the set of lexical competitors. I tested this claim with an examination of assimilation in first mentions of landmark names that did and did not have close competitors arising from the change in place feature of the word-final nasal. The result offered no support for the distinctiveness principle. A second prediction with respect to the incidence of schwa syncope in WS polysyllables was made but not tested because of a lack of available data. It was predicted that where the deletion of schwa leads to phonotactically acceptable onsets, speakers should refrain from hypo-articulating when introducing the word into the discourse for the first time to avoid activation of inappropriate competitors. It would be desirable to use additional material from the HCRC Map Task Corpus to supplement the few tokens already analysed and test this hypothesis. Ideally word identification responses should be elicited for the WS polysyllables to ascertain whether subjects respond to these stimuli with Real Word Responses that start with Strong initial syllables related to the resyllabified onset.

The analysis of /d/-deletion could also be extended to test the claim that -ED affixed words, such as *pebbled*, and *carved*, should be more predictable (because of their shared stem) and therefore undergo greater reduction than word-final /d/ segments that belong to the stem itself. The analysis in Chapter 10 suggests that words like *round* and *gold* are as likely to undergo reduction as *poisoned* and *reclaimed*. However the number of items in the analysis is small and would benefit from supplementary data.

11.5 Conclusion

This thesis argues that Lindblom's H & H theory ought to be tested on data from natural unscripted speech involving pairs of speakers engaged in a genuine communicative task. The task of communication is essentially one of message transmission rather than phonetic analysis, and, as Lindblom correctly observes, the high level of redundancy in linguistic form allows speakers to hypo-articulate

without their listener failing to understand. However, doubt is cast upon the cooperative and listener-oriented image of the speaker as presented by Lindblom. This thesis argues in favour of a more egocentric speaker who reduces articulatory effort without consideration for the effects it may have on her listener. In addition, Lindblom's distinctiveness constraint is rejected on the grounds that speakers appear to hypo-articulate beyond the point of lexical uniqueness.

References

Abercrombie, D. (1967) *Elements of General Phonetics.* Edinburgh: Edinburgh University Press.

Anderson, A. H., M. Bader, E. G. Bard, E. Boyle, G. M. Doherty-Sneddon, S. Garrod, S. Isard, J. C. Kowtko, J. M. McAllister, J. E. Miller, C. F. Sotillo, H. S. Thompson, and R. Weinert (1991) The HCRC Map Task Corpus. *Language and Speech*, **34**, 4, 351–366.

Anderson, A. H., E. G. Bard, C. F. Sotillo, A. Newlands, and G. M. Doherty-Sneddon (1997) Limited visual control of the intelligibility of speech in face-to-face dialogue. *Perception and Psychophysics*, **59**, 4, 580–592.

Anderson, A. H., S. C. Garrod, and A. J. Sanford (1983) The accessibility of pronominal antecedents as a function of episode shifts in narrative text. *Quarterly Journal of Experimental Psychology*, **35A**, 427–440.

Andrews, S. (1989) Frequency and neighborhood effects on lexical access: Activation or search? *Journal of Experimental Psychology: Human Perception and Performance*, **15**, 5, 802–814.

Andrews, S. (1992) Frequency and neighborhood effects on lexical access: Lexical similarity or orthographic redundancy? *Journal of Experimental Psychology: Learning, Memory and Cognition*, **18**, 2, 234–254.

Archangeli, D. (1988) Aspects of underspecification theory. *Phonology*, **5**, 183–207.

Archangeli, D. and D. Pulleyblank (1994) *Grounded Phonology.* Cambridge, MA: MIT Press.

Ariel, M. (1990) *Accessing Noun-Phrase Antecedents.* London: Routledge.

Ariel, M. (1991) The function of accessibility in a theory of grammar. *Journal of Pragmatics*, **16**, 443–463.

Baayen, R. H., R. Piepenbrock, and L. Gulikers (1995) *The CELEX Lexical Database (CD-ROM).* Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania. Version 2.5.

Bard, E. G. and A. H. Anderson (1983) The unintelligibility of speech to children. *Journal of Child Language*, **10**, 265–292.

Bard, E. G. and A. H. Anderson (1994) The unintelligibility of speech to children: Effects of referent availability. *Journal of Child Language*, **21**, 623–648.

Bard, E. G., L. Cooper, J. Kowtko, and C. Brew (1991) Psycholinguistic studies on incremental recognition of speech: A revised and extended introduction to the messy and the sticky. *Tech. Rep. DYANA Deliverable R1.3B*, University of Edinburgh.

Bard, E. G., A. J. Lowe, and G. T. M. Altmann (1989) The effect of repetition on words in recorded dictations. In *Proceedings of Eurospeech 1989, 1st European Conference on Speech Communication and Technology*, vol. 2, pp. 573–576.

Bard, E. G., R. C. Shillcock, and G. T. M. Altmann (1988) The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context. *Perception and Psychophysics*, **44**, 5, 395–408.

Bard, E. G., C. F. Sotillo, A. H. Anderson, G. M. Doherty-Sneddon, and A. Newlands (1995) The control of intelligibility in running speech. In *Proceedings of the XIIIth International Congress of Phonetic Sciences*, vol. 4, pp. 188–191.

Barry, M. C. (1985) A palatographic study of connected speech processes. *Cambridge Papers in Phonetics and Experimental Linguistics*, **4**, B1–B16.

Bates, S. A. R. (1995) *Towards a Definition of Schwa: an Acoustic Investigation of Vowel Reduction in English.* Ph.D. thesis, University of Edinburgh.

van Bergem, D. R. and F. J. Koopmans-van Beinum (1989) Vowel reduction in natural speech. In *Proceedings of Eurospeech 1989, 1st European Conference on Speech Communication and Technology*, vol. 2, pp. 285–288.

Berman, A. and M. Szamosi (1972) Observations on sentential stress. *Language*, **48**, 304–325.

Bock, J. K. and J. R. Mazzella (1983) Intonational marking of given and new information: Some consequences for comprehension. *Memory and Cognition*, **11**, 1, 64–76.

Bolinger, D. (1978) Intonation across languages. In J. H. Greenberg (ed.), *Universals of Human Language, Vol. 2: Phonology*, pp. 471–524. Stanford, CA: Stanford University Press.

Broe, M. (1993) The group-delay spectrogram: a tool for acoustic analysis.

Browman, C. P. and L. Goldstein (1989) Articulatory gestures as phonological units. *Phonology*, **6**, 2, 201–251.

Browman, C. P. and L. Goldstein (1990) Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston and M. E. Beckman (eds.), *Between the Grammar and Physics of Speech: Papers in Laboratory Phonology I*, pp. 341–376. Cambridge: Cambridge University Press.

Brown, G. (1977) *Listening to Spoken English.* London: Longman.

Brown, G. (1983) Prosodic structure and the given/new distinction. In A. Cutler and D. R. Ladd (eds.), *Prosody: Models and Measurements*, pp. 67–77. Berlin: Springer-Verlag.

Brown, G., A. H. Anderson, R. Shillcock, and G. Yule (1984) *Teaching Talk.* Cambridge: Cambridge University Press.

Brown, G., K. L. Currie, and J. Kenworthy (1980) *Questions of Intonation.* London: Croom Helm.

Brown, G. and G. Yule (1983) *Discourse Analysis.* Cambridge: Cambridge University Press.

Campbell, W. N. and S. D. Isard (1991) Segment durations in a syllable frame. *Journal of Phonetics*, **19**, 37–47.

Chafe, W. L. (1976) Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In C. N. Li (ed.), *Subject and Topic*, pp. 25–55. New York: Academic Press.

Chomsky, N. and M. Halle (1968) *The Sound Pattern of English.* New York: Harper and Row.

Clancy, P. M. (1980) Referential choice in English and Japanese narrative discourse. In W. L. Chafe (ed.), *The Pear Stories: cognitive, cultural, and linguistic aspects of narrative production*, Advances in discourse processes, Volume 3, pp. 127–202. Norwood, N.J.: Ablex Publishing Corporation.

Clark, H. H. (1992) *Arenas of Language Use.* The University of Chicago Press.

Clark, H. H. (1996) *Using Language.* Cambridge: Cambridge University Press.

Clark, H. H. and S. E. Brennan (1991) Grounding in communication. In L. Resnick, J. Levine, and S. Teasley (eds.), *Perspectives on Socially Shared Cognition*, pp. 127–149. Addison-Wesley.

Clark, H. H. and S. E. Haviland (1977) Comprehension and the Given-New contract. In R. O. Freedle (ed.), *Discourse Production and Comprehension*, Discourse processes: advances in research and theory, Volume 1, pp. 1–40. Norwood, N.J.: Ablex Publishing Corporation.

Clark, H. H. and E. F. Schaefer (1987a) Collaborating on contributions to conversations. *Language and Cognitive Processes*, **2**, 19–41.

Clark, H. H. and E. F. Schaefer (1987b) Concealing one's meaning from overhearers. *Journal of Memory and Language*, **26**, 209–225.

Clark, H. H. and C. J. Sengul (1979) In search of referents for nouns and pronouns. *Memory and Cognition*, **7**, 1, 35–41.

Clements, G. N. (1985) The geometry of phonological features. *Phonology Yearbook*, **2**, 225–252.

Clements, G. N. and S. J. Keyser (1983) *CV Phonology.* Cambridge, MA: MIT Press.

Cohen, P. S. and R. L. Mercer (1975) The phonological component of an automatic speech-recognition system. In D. R. Reddy (ed.), *Speech Recognition: Invited Papers Presented at the 1974 IEEE Symposium*, pp. 290–308. New York: Academic Press.

Cohn, A. C. (1993) Nasalisation in English: Phonology or phonetics. *Phonology*, **10**, 43–81.

Connine, C. M., D. G. Blasko, and D. Titone (1993) Do the beginnings of spoken words have a special status in auditory word recognition? *Journal of Memory and Language*, **32**, 193–210.

Cooper, W. E. and M. Danly (1981) Segmental and temporal aspects of utterance-final lengthening. *Phonetica*, **38**, 106–115.

Cooper, W. E. and J. Paccia-Cooper (1980) *Syntax and Speech.* Cambridge, MA: Harvard University Press.

Cooper, W. E., C. Soares, and A. Ham (1983) The influence of inter- and intra-speaker tempo differences on fundamental frequency and palatalisation. *The Journal of the Acoustical Society of America*, **73**, 5, 1723–1730.

Craig, C. H. and B. W. Kim (1990) Effects of time gating and word-length on isolated word recognition performance. *Journal of Speech and Hearing Research*, **33**, 4, 808–815.

Cutler, A. (1976) Phoneme-monitoring reaction time as a function of preceding intonation contour. *Perception and Psychophysics*, **20**, 55–60.

Cutler, A. and D. R. Ladd (1983) Comparative notes on terms and topics in the contributions. In A. Cutler and D. R. Ladd (eds.), *Prosody: Models and Measurements*, pp. 141–146. Berlin: Springer-Verlag.

Cutler, A. and D. Norris (1988) The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, **14**, 1, 113–121.

Dalby, J. M. (1984) *Phonetic Structure of Fast Speech in American English.* Ph.D. thesis, Indiana University.

Davies, B. L. (1997) *An Empirical Examination of Cooperation, Effort and Risk in Task-Oriented Dialogues.* Ph.D. thesis, University of Edinburgh.

Delattre, P. (1968) From acoustic cues to distinctive features. *Phonetica*, **18**, 198–230.

Eefting, W. (1991) The effect of "information value" and "accentuation" on the duration of Dutch words, syllables and segments. *The Journal of the Acoustical Society of America*, **89**, 1, 412–424.

Ellis, L. and W. J. Hardcastle (1997) An EPG study of alveolar to velar coarticulation in fast and careful speech: some preliminary observations. Internal document produced at Queen Margaret College, Edinburgh.

Fant, G. (1970) *Acoustic Theory of Speech Production.* The Hague: Mouton, 2nd edn.

Fisher, C. and H. Tokura (1995) The Given-New contract in speech to infants. *Journal of Memory and Language*, **34**, 287–310.

Forster, K. I. (1976) Accessing the mental lexicon. In R. J. Wales and E. Walker (eds.), *New Approaches to Language Mechanisms*, pp. 257–287. Amsterdam: North-Holland.

Fowler, C. A. (1980) Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, **8**, 113–133.

Fowler, C. A. (1988) Differential shortening of repeated content words produced in various communicative contexts. *Language and Speech*, **31**, 4, 307–319.

Fowler, C. A. and J. Housum (1987) Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language*, **26**, 489–504.

Francis, W. N. and H. Kucera (1982) *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston, MA: Houghton Mifflin. Assisted by Andrew W. Mackie.

Frauenfelder, U. H. and A. Lahiri (1989) Understanding words and word recognition: Can phonology help? In W. D. Marslen-Wilson (ed.), *Lexical Representation and Process*, pp. 319–341. Cambridge, MA: MIT Press.

Frauenfelder, U. H. and L. K. Tyler (1987) The process of spoken word recognition: An introduction. In U. H. Frauenfelder and L. K. Tyler (eds.), *Spoken Word Recognition*, Cognition Special Issue, volume 25, pp. 1–20. Cambridge, MA: MIT Press.

Frazier, L. (1987) Structure in auditory word recognition. *Cognition*, **25**, 157–187.

Fujimura, O. (1962) Analysis of nasal consonants. *The Journal of the Acoustical Society of America*, **34**, 12, 1865–1875.

Garvin, P. L. and P. Ladefoged (1963) Speaker identification and message identification in speech recognition. *Phonetica*, **9**, 193–199.

Gaskell, M. G. and W. D. Marslen-Wilson (1995) Modeling the perception of spoken words. In J. D. Moore and J. F. Lehman (eds.), *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society*, pp. 19–24. Mahwah, N.J.: Lawrence Erlbaum Associates.

Gaskell, M. G. and W. D. Marslen-Wilson (1996) Phonological variation and inference in lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, **22**, 1, 144–158.

Gay, T. (1978) Effect of speaking rate on vowel formant movements. *Journal of the Acoustical Society of America*, **63**, 1, 223–230.

Gernsbacher, M. A. and T. T. Givón (1995) Introduction: Coherence as a mental entity. In M. A. Gernsbacher and T. Givón (eds.), *Coherence in Spontaneous Text*, pp. vii–x. Amsterdam: John Benjamins Publishing Company.

Gimson, A. C. (1980) *An Introduction to the Pronunciation of English.* London: Edward Arnold, 2nd edn.

Givón, T. (1995) Coherence in text vs. coherence in mind. In M. A. Gernsbacher and T. Givón (eds.), *Coherence in Spontaneous Text*, pp. 59–115. Amsterdam: John Benjamins Publishing Company.

Glass, J. R. and V. W. Zue (1986) Detection and recognition of nasal consonants in American English. In *Proceedings of IEEE ICASSP-86*, pp. 2767–2770.

Goldsmith, J. A. (1976) *Autosegmental Phonology.* Ph.D. thesis, MIT.

Granit, R. (1977) *The Purposive Brain.* Cambridge, MA: MIT Press.

Green, K. P. and L. W. Norrix (1997) Acoustic cues to place of articulation and the McGurk effect: The role of release bursts, aspiration, and formant transitions. *Journal of Speech, Language, and Hearing Research*, **40**, 3, 646–665.

Grice, H. P. (1975) Logic and conversation. In P. Cole and J. L. Morgan (eds.), *Syntax and Semantics, Vol 3: Speech Acts*, pp. 41–58. New York: Academic Press.

Grosjean, F. (1980) Spoken word recognition processes and the gating paradigm. *Perception and Psychophysics*, **28**, 267–283.

Grosjean, F. (1985) The recognition of words after their acoustic offset: Evidence and implications. *Perception and Psychophysics*, **38**, 4, 299–310.

Grosz, B. and C. Sidner (1986) Attention, intention, and the structure of discourse. *Computational Linguistics*, **12**, 175–206.

Gundel, J., N. Hedberg, and R. Zacharski (1993) Cognitive status and the form of referring expressions in discourse. *Language*, **69**, 274–307.

Hajičová, E. (1991) Topic-focus articulation and coreference in models of discourse production. *Journal of Pragmatics*, **16**, 157–166.

Halliday, M. A. K. (1967) Notes on transitivity and theme in English, Part 2. *Journal of Linguistics*, **3**, 199–244.

Halliday, M. A. K. (1992) *Spoken and Written Language.* Oxford: Oxford University Press, third impression edn.

Harris, M. S. and N. Umeda (1974) Effect of speaking mode on temporal factors in speech: Vowel duration. *The Journal of the Acoustical Society of America*, **56**, 3, 1016–1018.

Hattori, S., K. Yamamoto, and O. Fujimura (1958) Nasalization of vowels in relation to nasals. *The Journal of the Acoustical Society of America*, **30**, 4, 267–274.

Haviland, S. and H. Clark (1974) What's New? Acquiring New information as a process in comprehension. *Journal of Verbal Learning and Verbal Behaviour*, **13**, 512–521.

Hawkins, S. and P. Warren (1994) Phonetic influences on the intelligibility of conversational speech. *Journal of Phonetics*, **22**, 493–511.

Hayes, B. (1992) Commentary on F. Nolan, "The descriptive role of segments: Evidence from assimilation". In G. J. Docherty and D. R. Ladd (eds.), *Gesture, Segment, Prosody: Papers in Laboratory Phonology II*, pp. 280–286. Cambridge: Cambridge University Press.

Heeman, P. A. and G. Hirst (1995) Collaborating on referring expressions. *Computational Linguistics*, **21**, 3, 351–382.

Heffner, R.-M. S. (1960) *General Phonetics.* The University of Wisconsin Press.

Hieronymus, J., M. Alexander, C. Bennett, I. Cohen, D. Davies, J. Dalby, J. Laver, W. Barry, A. Fourcin, and J. Wells (1990) Proposed speech segmentation criteria for the SCRIBE project. *Tech. rep.*, Centre for Speech Technology Research, University of Edinburgh and Department of Phonetics and Linguistics, University College London. Preliminary draft, Rev.2.

Holst, T. and F. J. Nolan (1995) The influence of syntactic structures on [s] to [ʃ] assimilation. In B. Connell and A. Arvaniti (eds.), *Phonology and Phonetic Evidence: Papers in Laboratory Phonology IV*, pp. 315–333. Cambridge: Cambridge University Press.

Horne, M. (1991) Why do speakers accent "Given" information? In *Proceedings of Eurospeech 1991, 2nd European Conference on Speech Communication and Technology*, vol. 3, pp. 1279–1282.

House, A. S. (1957) Analog studies of nasal consonants. *Journal of Speech and Hearing Disorders*, **22**, 2, 190–204.

Hunnicut, S. (1985) Intelligibility versus redundancy – conditions of dependency. *Language and Speech*, **28**, 45–56.

Ingram, J. and T. Mylne (1994) Perceptual parsing of nasal vowels. In *Proceedings of the 1994 International Conference on Spoken Language Processing*, vol. 2, pp. 495–498. Yokohama: Acoustical Society of Japan.

Johnson-Laird, P. N. (1983) *Mental Models.* Cambridge: Cambridge University Press.

Johnson-Laird, P. N. (1987) The mental representation of the meaning of words. In U. H. Frauenfelder and L. K. Tyler (eds.), *Spoken Word Recognition*, Cognition Special Issue, volume 25, pp. 189–211. Cambridge, MA: MIT Press.

Kelly, M. L. (1993) *Lexical Segmentation and Word Recognition in Fluent Aphasia.* Ph.D. thesis, University of Edinburgh.

Kelso, J. A. S., E. L. Saltzman, and B. Tuller (1986) The dynamical perspective on speech production: data and theory. *Journal of Phonetics*, **14**, 29–59.

Kerswill, P. E. (1985) A sociophonetic study of connected speech processes in Cambridge English: An outline and some results. *Cambridge Papers in Phonetics and Experimental Linguistics*, **4**, 1–39.

Kiparsky, P. (1982) Lexical morphology and phonology. In L. S. of Korea (ed.), *Linguistics in the Morning Calm: Selected Papers from SICOL-1981*, pp. 3–92. Seoul, Korea: Hanshin Publishing Company.

Kiparsky, P. (1986) Comment on W. Labov's paper (chapter 19). In J. S. Perkell and D. H. Klatt (eds.), *Invariance and Variability in Speech Processes*, pp. 423–424. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Kitazawa, S. and S. Doshita (1984) Nasal consonant discrimination by vowel independent features. *Studia Phonologica*, **XVIII**, 46–58.

Klatt, D. H. (1986) The problem of variability in speech recognition and in models of speech perception. In J. S. Perkell and D. H. Klatt (eds.), *Invariance and Variability in Speech Processes*, pp. 300–319. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Klatt, D. H. (1989) Review of selected models of speech perception. In W. D. Marslen-Wilson (ed.), *Lexical Representation and Process*, pp. 169–226. Cambridge, MA: MIT Press.

Kohler, K. L. (1990) Segmental reduction in connected speech in German: Phonological facts and phonetic explanations. In W. J. Hardcastle and A. Marchal (eds.), *Speech Production and Speech Modelling*, pp. 69–92. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Kowtko, J. C. (1996) *The Function of Intonation in Task Oriented Dialogue.* Ph.D. thesis, University of Edinburgh.

Kuehn, D. P. and K. L. Moll (1976) A cineradiographic study of VC and CV articulatory velocities. *Journal of Phonetics*, **4**, 303–320.

Kuno, S. (1974) Lexical and contextual meaning. *Linguistic Inquiry*, **5**, 469–477.

Kuno, S. (1978) Generative discourse analysis in America. In W. Dressler (ed.), *Current Trends in Textlinguistics*, pp. 275–294. Berlin and New York: de Gruyter.

Kurowski, K. and S. E. Blumstein (1984) Perceptual integration of the murmur and formant transitions for place of articulation in nasal consonants. *The Journal of the Acoustical Society of America*, **76**, 2, 383–390.

Kurowski, K. and S. E. Blumstein (1987) Acoustic properties for place of articulation in nasal consonants. *The Journal of the Acoustical Society of America*, **81**, 6, 1917–1927.

Ladd, D. R. (1996) *Intonational Phonology.* Cambridge: Cambridge University Press.

Ladefoged, P. (1982) *A Course in Phonetics.* New York: Harcourt Brace Jovanovich, 2nd edn.

Ladefoged, P. and I. Maddieson (1989) Multiply articulated segments and the feature hierarchy. *UCLA Working Papers in Phonetics*, **72**, 116–138.

Lahiri, A. and W. D. Marslen-Wilson (1991) The mental representation of lexical form: A phonological approach to the recognition lexicon. *Cognition*, **38**, 245–294.

Lambrecht, K. (1994) *Information Structure and Sentence Form: topic, focus, and the mental representations of discourse referents.* Cambridge: Cambridge University Press.

Larkey, L. S., J. Wald, and W. Strange (1978) Perception of synthetic nasal consonants in initial and final syllable position. *Perception and Psychophysics*, **23**, 299–312.

Laver, J., C. Bennett, I. Cohen, J. Dalby, D. Davies, S. Hiller, M. McAllister, and A. Sutherland (1989a) ATR/CSTR Speech Database Project. *Tech. Rep. 2*, Centre for Speech Technology Research, University of Edinburgh.

Laver, J., C. Bennett, I. Cohen, J. Dalby, D. Davies, and M. McAllister (1988) ATR/CSTR Speech Database Project. *Tech. Rep. 1*, Centre for Speech Technology Research, University of Edinburgh.

Laver, J., J. Hieronymus, C. Bennett, I. Cohen, J. Dalby, D. Davies, S. Hiller, M. McAllister, and E. M. Purves (1989b) ATR/CSTR Speech Database Project. *Tech. Rep. 3*, Centre for Speech Technology Research, University of Edinburgh.

Levelt, W. J. M. (1989) *Speaking: from intention to articulation.* Cambridge, MA: MIT Press.

Liberman, A. M., P. C. Delattre, F. S. Cooper, and L. J. Gerstman (1954) The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monograph*, **68**, 1–13.

Lieberman, P. (1963) Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, **6**, 172–187.

Lindblom, B. (1963) Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, **35**, 11, 1773–1781.

Lindblom, B. (1983a) Economy of speech gestures. In P. F. MacNeilage (ed.), *The Production of Speech*, pp. 217–245. Heidelberg: Springer-Verlag.

Lindblom, B. (1983b) On the teleological nature of speech processes. *Speech Communication*, **2**, 155–158.

Lindblom, B. (1990a) Explaining phonetic variation: a sketch of the H & H theory. In W. J. Hardcastle and A. Marchal (eds.), *Speech Production and Speech Modelling*, pp. 403–439. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Lindblom, B. (1990b) On the notion of "possible speech sound". *Journal of Phonetics*, **18**, 135–152.

Lindblom, B., S. Brownlee, B. Davis, and S.-J. Moon (1992) Speech transforms. *Speech Communication*, **11**, 357–368.

Lindblom, B. and O. Engstrand (1989) In what sense is speech quantal? *Journal of Phonetics*, **17**, 107–121.

Lindblom, B., J. Lubker, and T. Gay (1979) Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation. *Journal of Phonetics*, **7**, 147–161.

Lindblom, B. and P. MacNeilage (1986) Action theory: Problems and alternative approaches. *Journal of Phonetics*, **14**, 117–132.

Lindblom, B., P. NacNeilage, and M. Studdert-Kennedy (1983) Self-organizing processes and the explanation of phonological universals. *Linguistics*, **21**, 1, 181–203.

Lindblom, B., J. Perkell, and D. Klatt (1986) Preface. In J. S. Perkell and D. H. Klatt (eds.), *Invariance and Variability in Speech Processes*, pp. i–iv. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Lindblom, B. and J. Sundberg (1971) Acoustical consequences of lip, tongue, jaw and larynx movement. *Journal of the Acoustical Society of America*, **50**, 4, 1166–1179.

Lindqvist-Gauffin, J. and J. Sundberg (1976) Acoustic properties of the nasal tract. *Phonetica*, **33**, 161–168.

Lively, S. E., D. B. Pisoni, and S. D. Goldinger (1994) Spoken word recognition: Research and theory. In M. A. Gernsbacher (ed.), *Handbook of Psycholinguistics*, pp. 265–301. San Diego, CA: Academic Press.

Luce, P. A. (1986) *Neighborhoods of Words in the Mental Lexicon.* Ph.D. thesis, Indiana University, Bloomington.

Luce, P. A., D. B. Pisoni, and S. D. Goldinger (1990) Similarity neighborhoods of spoken words. In G. T. M. Altmann (ed.), *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives*, pp. 122–147. Cambridge, MA: MIT Press.

Luce, R. D. (1959) *Individual Choice Behavior.* New York: Wiley.

Malécot, A. (1956) Acoustic cues for nasal consonants: An experimental study involving a tape-splicing technique. *Language*, **32**, 274–284.

Marslen-Wilson, W. D. (1984) Function and process in spoken word-recognition. In H. Bouma and D. G. Bouwhuis (eds.), *Attention and Performance X: Control of Language Processes*, pp. 125–150. Hillsdale, N.J.: Erlbaum.

Marslen-Wilson, W. D. (1987) Functional parallelism in spoken word-recognition. In U. H. Frauenfelder and L. K. Tyler (eds.), *Spoken Word Recognition*, Cognition Special Issue, volume 25, pp. 71–102. Cambridge, MA: MIT Press.

Marslen-Wilson, W. D. (1989) Access and integration: Projecting sound onto meaning. In W. D. Marslen-Wilson (ed.), *Lexical Representation and Process*, pp. 3–24. Cambridge, MA: MIT Press.

Marslen-Wilson, W. D. (1993) Issues of process and representation in lexical access. In G. T. M. Altmann and R. C. Shillcock (eds.), *Cognitive Models of Speech Processing: The Second Sperlonga Meeting*, pp. 187–210. Hove, UK: Lawrence Erlbaum Associates.

Marslen-Wilson, W. D. and M. G. Gaskell (1992) Match and mismatch in lexical access [abstract]. *International Journal of Psychology*, **27**, 3, 61.

Marslen-Wilson, W. D., H. E. Moss, and S. van Halen (1996) Perceptual distance and competition in lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, **22**, 6, 1376–1392.

Marslen-Wilson, W. D., A. Nix, and G. Gaskell (1995) Phonological variation in lexical access: Abstractness, inference and English place assimilation. *Language and Cognitive Processes*, **10**, 3/4, 285–308.

Marslen-Wilson, W. D. and L. K. Tyler (1980) The temporal structure of spoken language understanding. *Cognition*, **8**, 1–71.

Marslen-Wilson, W. D. and P. Warren (1994) Levels of perceptual representation and process in lexical access: Words, phonemes and features. *Psychological Review*, **101**, 653–675.

Marslen-Wilson, W. D. and A. Welsh (1978) Processing interactions during word-recognition in continuous speech. *Cognitive Psychology*, **10**, 29–63.

Marslen-Wilson, W. D. and P. Zwitserlood (1989) Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance*, **15**, 3, 576–585.

Mayo, C., M. P. Aylett, and D. R. Ladd (1997) Prosodic transcription of Glasgow English: an evaluation study of GlaToBI. In *Intonation: Theory, Models and Applications; Proceedings of an ESCA Workshop*, pp. 231–234.

McAllister, J., C. F. Sotillo, and E. G. Bard (1991) The effect of addressee familiarity on word duration. In *Proceedings of the XIIth International Congress of Phonetic Sciences*, vol. 4, pp. 426–429.

McAllister, J. M. (1988) The use of context in auditory word recognition. *Perception and Psychophysics*, **44**, 1, 94–97.

McAllister, J. M. (1989) *Lexical Stress and Lexical Access: Effects in Read and Spontaneous Speech.* Ph.D. thesis, University of Edinburgh.

McAllister, J. M., C. F. Sotillo, E. G. Bard, and A. H. Anderson (1990) Using the Map Task to investigate variability in speech. *Occasional paper*, Department of Linguistics, University of Edinburgh.

McClelland, J. L. and J. L. Elman (1986) The TRACE model of speech perception. *Cognitive Psychology*, **18**, 1–86.

McClelland, J. L. and D. E. Rumelhart (1981) An interactive activation model of context effects in letter perception: Part 1. An account of the basic findings. *Psychological Review*, **88**, 375–407.

McClelland, J. L. and D. E. Rumelhart (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* Cambridge, MA: Bradford Books.

Mehta, G. and A. Cutler (1988) Detection of target phonemes in spontaneous and read speech. *Language and Speech*, **31**, 2, 135–156.

Miller, G. A. and P. E. Nicely (1955) An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America*, **27**, 338–352.

Molloy, L. (1997) On the use of durational plausibility in speech recognition. Talk presented to the Edinburgh University Phonetics–Phonology Workshop/CSTR Seminar Series.

Monsell, S. (1985) Repetition and the lexicon. In A. W. Ellis (ed.), *Progress in the Psychology of Language*, vol. 2, pp. 147–195. London: Lawrence Erlbaum Associates Ltd.

Moon, S.-J. and B. Lindblom (1989) Formant undershoot in clear and citation-form speech: a second progress report. *Speech Transmission Laboratory – Quarterly Progress and Status Report*, **1**, 121–123.

Moss, H. E. and W. D. Marslen-Wilson (1993) Access to word meanings during spoken language comprehension: Effects of sentential semantic context. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **19**, 6, 1254–1276.

Nolan, F. J. (1983) *The Phonetic Bases of Speaker Recognition.* Cambridge: Cambridge University Press.

Nolan, F. J. (1992) The descriptive role of segments: Evidence from assimilation. In G. J. Docherty and D. R. Ladd (eds.), *Gesture, Segment, Prosody: Papers in Laboratory Phonology II*, pp. 261–289. Cambridge: Cambridge University Press.

Nolan, F. J., T. Holst, and B. Kuhnert (1996) Modeling [s] to [ʃ] accomodation in English. *Journal of Phonetics*, **24**, 1, 113–137.

Nooteboom, S. G. (1981) Lexical retrieval from fragments of spoken words: Beginnings vs endings. *Journal of Phonetics*, **9**, 407–424.

Nooteboom, S. G. and J. G. Kruyt (1987) Accents, focus distribution, and the perceived distribution of given and new information: An experiment. *The Journal of the Acoustical Society of America*, **82**, 1512–1524.

Öhman, S. E. G. (1966) Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America*, **39**, 1, 151–168.

Oller, D. K. (1973) The effect of position in utterance on speech segment duration in English. *The Journal of the Acoustical Society of America*, **54**, 5, 1235–1247.

Oshika, B. T., V. W. Zue, R. V. Weeks, H. Neu, and J. Aurbach (1975) The role of phonological rules in speech understanding research. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **ASSP-23**, 1, 104–112.

Paradis, C. and J.-F. Prunet (eds.) (1991) *Phonetics and Phonology II: The Special Status of Coronals.* San Diego, CA: Academic Press.

Peterson, G. E. and I. Lehiste (1960) Duration of syllable nuclei in English. *The Journal of the Acoustical Society of America*, **32**, 6, 693–703.

Pickett, J. M. (1965) Some acoustic cues for synthesis of the /n-d/ distinction. *The Journal of the Acoustical Society of America*, **38**, 474–477.

Pierrehumbert, J. B. (1980) *The Phonology and Phonetics of English Intonation.* Ph.D. thesis, MIT.

Prince, E. F. (1981) Toward a taxonomy of Given-New information. In P. Cole (ed.), *Radical Pragmatics*, pp. 223–255. London: Academic Press.

Qi, Y. and R. A. Fox (1992) Analysis of nasal consonants using perceptual linear prediction. *The Journal of the Acoustical Society of America*, **91**, 3, 1718–1726.

Rakerd, B., W. Sennett, and C. A. Fowler (1987) Domain-final lengthening and foot-level shortening in spoken English. *Phonetica*, **44**, 147–155.

Recasens, D. (1983) Place cues for nasal consonants with special reference to Catalan. *The Journal of the Acoustical Society of America*, **73**, 1346–1353.

Repp, B. H. (1986) Perception of the [m]-[n] distinction in CV syllables. *The Journal of the Acoustical Society of America*, **79**, 6, 1987–1999.

Rietveld, A. C. M. and F. J. Koopmans-van Beinum (1987) Vowel reduction and stress. *Speech Communication*, **6**, 217–229.

Rooney, E. J. (1990) *Nasality in Automatic Speaker Verification.* Ph.D. thesis, University of Edinburgh.

Sagey, E. (1986) *The Representation of Features and Relations in Non-Linear Phonology.* Ph.D. thesis, MIT.

Samuel, A. G. (1981) Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General*, **110**, 474–494.

Samuel, A. G. (1987) Lexical uniqueness effects on phonemic restoration. *Journal of Memory and Language*, **26**, 36–56.

van Santen, J. P. H. and J. P. Olive (1990) The analysis of contextual effects on segmental duration. *Computer Speech and Language*, **4**, 359–390.

Savin, H. B. (1963) Word-frequency effect and errors in the perception of speech. *The Journal of the Acoustical Society of America*, **35**, 200–206.

Schwartz, C., G. Davidson, A. Seaton, and V. Tebbit (eds.) (1988) *Chambers English Dictionary.* W & R Chambers Ltd and Cambridge University Press, 7th edn.

Shearme, J. N. and J. N. Holmes (1961) An experimental study of the classification of sounds in continuous speech according to their distribution in the formant 1 – formant 2 plane. In *Proceedings of the IVth International Congress of Phonetic Sciences*, pp. 232–240. Helsinki.

Shields, L. W. and D. A. Balota (1991) Repetition and associative context effects in speech production. *Language and Speech*, **34**, 1, 47–55.

Shillcock, R. C., J. Hicks, P. Cairns, N. Chater, and J. P. Levy (1996) Phonological reduction, assimilation, intra-word information structure, and the evolution of the lexicon in English: Why fast speech isn't confusing. In *Proceedings of the 18th Annual Cognitive Science Society*, pp. 233–238.

Shockey, L. (1974) Phonetic and phonological properties of connected speech. *Ohio State Working Papers in Linguistics*, **17**, iv–143.

Shockey, L. and Z. S. Bond (1980) Phonological processes in speech addressed to children. *Phonetica*, **37**, 267–274.

Smits, R., L. Tenbosch, and R. Collier (1996) Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants. 1: Perception experiment. *The Journal of the Acoustical Society of America*, **100**, 6, 3852–3864.

van Son, R. J. J. H. and L. C. W. Pols (1989) Comparing formant movements in fast and normal rate speech. In *Proceedings of Eurospeech 1989, 1st European Conference on Speech Communication and Technology*, vol. 2, pp. 665–668.

Steriade, D. (1987) Redundant values. In A. Bosch, B. Need, and E. Schiller (eds.), *Papers from the 23rd Annual Regional Meeting of the Chicago Linguistic Society, Part Two: Parasession on Autosegmental and Metrical Phonology*, pp. 339–362. Chicago: Chicago Linguistic Society.

Stevens, K. N. and A. S. House (1963) Perturbation of vowel articulations by consonantal context: an acoustical study. *Journal of Speech and Hearing Research*, **6**, 111–128.

Tabossi, P. (1993) Connections, competitions, and cohorts: Comments on the chapters by Marslen-Wilson; Norris; and Bard and Shillcock. In G. T. M. Altmann and R. C. Shillcock (eds.), *Cognitive Models of Speech Processing: The Second Sperlonga Meeting*, pp. 277–294. Hove, UK: Lawrence Erlbaum Associates.

Tannen, D. (ed.) (1984) *Coherence in Spoken and Written Discourse.* Advances in discourse processes, Volume 12. Norwood, N.J.: Ablex Publishing.

Terken, J. and S. G. Nooteboom (1987) Opposite effects of accentuation and deaccentuation on verification latencies for Given and New information. *Language and Cognitive Processes*, **2**, 3/4, 145–163.

Terken, J. M. B. (1985) *Use and Function of Accentuation: Some Experiments.* Ph.D. thesis, Leyden University.

Thorndike, E. L. and I. Lorge (1944) *The Teacher's Word Book of 30,000 Words.* New York: Teacher's College, Columbia University.

Tiffany, W. R. (1959) Nonrandom sources of variation in vowel quality. *Journal of Speech and Hearing Research*, **2**, 305–317.

Tinbergen, N. (1951) *The Study of Instinct.* Oxford: Clarendon Press.

Tuller, B., K. S. Harris, and J. A. S. Kelso (1982) Stress and rate: Differential transformations of articulation. *The Journal of the Acoustical Society of America*, **71**, 6, 1534–1543.

Tyler, L. K. (1984) The structure of the initial cohort: Evidence from gating. *Perception and Psychophysics*, **36**, 5, 417–427.

Tyler, L. K., R. Waksler, and W. D. Marslen-Wilson (1993) Representation and access of derived words in English. In G. T. M. Altmann and R. C. Shillcock

(eds.), *Cognitive Models of Speech Processing: The Second Sperlonga Meeting*, pp. 125–140. Hove, UK: Lawrence Erlbaum Associates.

Tyler, L. K. and J. Wessels (1983) Quantifying contextual contributions to word-recognition processes. *Perception and Psychophysics*, **36**, 217–222.

Umeda, N. (1975) Vowel duration in American English. *The Journal of the Acoustical Society of America*, **58**, 2, 434–445.

Vallduví, E. (to appear) Information packaging: A survey. *Tech. Rep. HCRC/RP-44*, Human Communication Research Centre, University of Edinburgh.

Vallduví, E. and R. Zacharski (1994) Accenting phenomena, association with focus, and the recursiveness of focus-ground. In P. Dekker and M. Stokhof (eds.), *Proceedings of the Ninth Amsterdam Colloquium*, pp. 683–702. Amsterdam: Institute for Logic, Language and Computation.

Walker, C. H. and F. R. Yekovich (1987) Activation and use of script-based antecedents in anaphoric reference. *Journal of Memory and Language*, **26**, 673–691.

Walker, M. A. (1992) When Given information is accented: Repetition, paraphrase and inference in dialogue. In *Proceedings of the Institute for Research in Cognitive Science Workshop on Prosody in Natural Speech: IRCS Report number 92-37*, pp. 231–240. University of Pennsylvania.

Walker, M. A. (1996) Limited attention and discourse structure. *Computational Linguistics*, **22**, 2, 255–264.

Warren, P. and W. D. Marslen-Wilson (1987) Continuous uptake of acoustic cues in spoken word-recognition. *Perception and Psychophysics*, **41**, 262–275.

Warren, P. and W. D. Marslen-Wilson (1988) Cues to lexical choice: Discriminating place and voice. *Perception and Psychophysics*, **43**, 21–30.

Wayland, S. C., A. Wingfield, and H. Goodglass (1989) Recognition of isolated words – the dynamics of cohort reduction. *Applied Psycholinguistics*, **10**, 4, 475–487.

Wightman, C. W., S. Shattuck-Hufnagel, M. Ostendorf, and P. J. Price (1992) Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America*, **91**, 3, 1707–1717.

Wilkes-Gibbs, D. (1995) Coherence in collaboration: Some examples from conversation. In M. A. Gernsbacher and T. Givón (eds.), *Coherence in Spontaneous Text*, pp. 239–267. Amsterdam: John Benjamins Publishing Company.

Wolf, J. J. (1972) Efficient acoustic parameters for speaker recognition. *The Journal of the Acoustical Society of America*, **51**, 2044–2056.

Yegnanarayana, B. (1981) Speech analysis by pole-zero decomposition of short-time spectra. *Signal Processing*, **3**, 5–17.

Zipf, G. K. (1935) *The Psychology of Language.* Houghton-Mifflin.

Ziv, Y. (1996) Review of Lambrecht's "Information structure and sentence form: Topic, focus and the mental representations of discourse referents". *Journal of Pragmatics*, **26**, 702–706.

Zwicky, A. M. (1972) Note on a phonological hierarchy in English. In R. P. Stockwell and R. K. S. Macaulay (eds.), *Linguistic Change and Generative Theory*, pp. 275–301. Bloomington: Indiana University Press.

Zwitserlood, P. (1989) The locus of the effects of the sentential-semantic context in spoken-word processing. *Cognition*, **32**, 25–64.

Appendix A

The HCRC Map Task Corpus: full design

LAYER ONE						Subject Quadruple			
						1	2	3	4
Dialog	**Fam**	**Giver**	**Follwr**	**Contr**	**Match**	**Reduction**			
1	−	a1	b1	+	+	1	4	3	2
2	−	b2	a2	+	−	2	1	4	3
3	+	a2	a1	−	+	3	2	1	4
4	+	b1	b2	−	−	4	3	2	1
5	−	a2	b2	−	+	3	2	1	4
6	−	b1	a1	−	−	4	3	2	1
7	+	a1	a2	+	+	1	4	3	2
8	+	b2	b1	+	−	2	1	4	3

LAYER TWO						Subject Quadruple			
						5	6	7	8
Dialog	**Fam**	**Giver**	**Follwr**	**Contr**	**Match**	**Reduction**			
1	+	a1	a2	+	+	1	4	3	2
2	+	b2	b1	+	−	2	1	4	3
3	−	a2	b2	−	+	3	2	1	4
4	−	b1	a1	−	−	4	3	2	1
5	+	a2	a1	−	+	3	2	1	4
6	+	b1	b2	−	−	4	3	2	1
7	−	a1	b1	+	+	1	4	3	2
8	−	b2	a2	+	−	2	1	4	3

Note on pair codes:

a1 and a2 are the two members of one familiar pair;
b1 and b2 are the 2 members of the other;
A subject QUAD comprises an a-pair and a b-pair,
where neither member of each pair knows either member
of the other.

Table A.1: Allocation of maps and subjects to conversations 1 to 8

Appendix B

IPA and SAM-PA correspondence tables

Example	IPA	SAM-PA
pit	ɪ	I
pet	ɛ	E
pat	a	{
putt	ʌ	V
pot	ɒ	Q
put	ʊ	U
another	ə	@
bean	i	i:
bath	ɑ	A:
bought	ɔ	O:
boon	u	u:
burn	ɜ	3:
bay	eɪ	eI
buy	aɪ	aI
boy	ɔɪ	OI
no	əʊ	@U
brow	aʊ	aU
peer	ɪə	I@
pair	ɛə	E@
poor	ʊə	U@
timbre	ɛ̃	{~
détente	ɑ̃ː	A~:
lingerie	æ̃ː	{~:
bouill**o**n	ɒ̃ː	O~:

Table B.1: Relation between IPA and SAM-PA transcription symbols: English vowels (RP)

Example	IPA	SAM-PA
pat	p	p
bat	b	b
tot	t	t
dot	d	d
cot	k	k
got	ɡ	g
mad	m	m
need	n	n
ba**ng**	ŋ	N
lad	l	l
rat	r	r
fat	f	f
vat	v	v
thin	θ	T
then	ð	D
sap	s	s
zap	z	z
sheep	ʃ	S
measu**re**	ʒ	Z
yank	j	j
lo**ch**	x	x
had	h	h
wipe	w	w
cheap	ʧ	tS
jeep	ʤ	dZ
idealis**m**	m̩	m,
burde**n**	n̩	n,
baco**n**	ŋ̩	N,
dangl**e**	l̩	l,
fathe**r**	*	r*
	(possible linking 'r')	

Table B.2: Relation between IPA and SAM-PA transcription symbols: English consonants

Appendix C

*U coding of the HCRC Map Task Corpus

C.1 Feature codes

These codes remain constant for any one feature and can be added automatically

```
CODE          MEANING

tdel          t-deletion              ie phonological
glott         glottalisation             modification/reduction
ddel          d-deletion                 category
nass          nasal assmilation
SW            polysyll-strong-initial
WS            polysyll-weak-initial
omo           odd-man-out

contrsh       contrast - shared
                eg both have east and west lake
contrnsh      contrast - not shared
                eg Giver has both, Follower only east
nocontrsh     no contrast - shared
                eg both only have east lake
nocontrnsh    no contrast - not shared
                eg Giver has only east, Follower both
These labels only attach to
   east/west lake          diamond/gold mine
  white/slate mountain       crane/green bay
```

CODE	MEANING
same	feature is shared by Giver and Follower
dif	feature is different on Giver's and Follower's maps:
f01	Giver has this feature, Follower does not
f10	Follower has this feature, Giver does not
feat	diff label, diff picture, same location ie different feature but in same position as partner's feature eg swamp and crocodiles feat01 = Giver's feature eg swamp feat10 = Follower's feature eg crocodiles
flab	diff label, same picture, same location ie name change flab01 = Giver has label with phonological reduction flab10 = Follower has label with phonological reduction
fnum	label(x) picture(x) location(i) label(x) picture(x) location(r) Giver has two features while Follower has one Only one is relevant to the path; Follower has irrelevant one fnum 01 = relevant feature to path (Giver only) fnum same = irrelevant feature to path (shared by G and F)
i	for irrelevant, attached to feature - eg picketjjfence i
r	for relevant, attached to feature - eg picketjjfence r

C.2 Utterance codes

These codes relate to the utterance itself in which reference to the feature is made.

a) codes relating to whole utterance

CODE	MEANING
in	instruction: IG telling IF to do something
loc	location: a statement about the location of something
pos	position: a statement about one's position
ex	existence : statement about the existence of a landmark eg You've got neither a graveyard nor a fast flowing river.
qu	question: any question except those listed below
qu loc	question location: a question about the location of something
qu ex	question existence : question about the existence of a landmark eg Do you have a graveyard? (NB note difference between this and use of "intro qu")
qcomply	questioning partner's compliance: eg "Have yo done that?", "OK?"
qunsp	unspecific question: eg "Eh?", "What?"
resp	response to a statement:
qresp	response to a question:
+	a POSITIVE response, eg "yes", "uh-huh"
-	a NEGATIVE response, eg "no", "nup"
deny	denial: explicit denial of a feature, eg "I haven't got a bridge"
ack	explicit acknowledgement of a feature, the converse of 'deny'
dir	direction: any expression of direction, eg up, down, towards
dis	distance: any expression of distance, even very vague. eg a wee bit, two inches

b) codes for landmark references

CODE	MEANING
intro	introduction of a feature -
intro men	introduction of a feature by mention eg "Go up to THE BRIDGE"
intro loc	introduction of a feature by mention with location eg "Go up to THE BRIDGE AT THE TOP OF THE PAGE"
intro qu	introduction of a feature by question eg "Do you have THE BRIDGE?"
intro qu loc	introduction of a feature by questionn with location eg "Do you have THE BRIDGE AT THE TOP OF THE PAGE?"
rep	repetition: repeated mention of a feature
def	definite: reference to a feature by definite NP
indef	indefinite: reference to a feature by indefinite NP
null	no article: reference to a feature with no article eg "Go to banana tree" "do you have gazelles?"
pro	pronoun: reference to a feature by a pronoun, eg "it"
el	ellipsis: reference to a feature by ellipsis
relpro	relative pronoun: reference to a feature by the rel pro "that"
dctc	deictic: reference to a feature using deixis
poss	possessive: use of possessive pronoun eg "my", "your"
dem	demonstrative: for "THIS tree" or "THAT mountain"
l	for literal use of label
nl	for non-literal term - eg "pile of stones" for "tor"
rl	for reduced literal term - eg "river" for "fast-flowing river"
d	for disfluent - eg "di-" for "diamond mine"

Appendix D

Subject Responses to Stimulus Words

D.1 Classification of response data, indicating how many Real Word Responses were available for analysis

Key to Table D.1:

- **target**: the stimulus word
- **items**: the number of different word tokens that were presented in the series of intelligibility experiments
- **N**: the total number of subject responses to the stimulus
- **dataset**: the phonological reduction category to which the stimulus belongs
- **1**: the number of perfect responses
- **1***: the number of responses where the word was obviously mispelled, either by doubling consonants where there should be one, e.g. *babboons*, or omitting a consonant where there should be two, e.g. *alotments*
- **1.**: the number of responses where a morphological affix was missing, either the word was singular instead of plural (*elephant, crocodile*), or the past tense *-ed* was absent e.g. *poison*
- **1+**: the number of responses with additional morphological material, e.g. the plural form was given to a singular stimulus e.g. *saloons*
- **1'**: the number of responses with a plural rather than *-ed* affix, used exclusively for the /d/-deletion items
- '–': the number of null responses, i.e. where no response of any kind was given
- **typo**: the number of responses with non-words that look like a poorly spelled version of the stimulus word
- **NW**: the number of Non-Word responses (i.e. not found in the CELEX database (Baayen *et al.*, 1995) or Chambers English Dictionary (Schwartz *et al.*, 1988))
- **RWR**: the number of Real Word responses (i.e. genuine lexical competitors)

target	items	N	dataset	1	1*	1.	1+	1'	'–'	typo	NW	RWR
abandoned	3	27	poly:WS	18	2	5	0	0	0	2	0	0
allotments	12	152	poly:WS	130	1	1	0	0	7	1	5	7
apache	7	63	poly:WS	42	4	0	0	0	5	8	2	2
baboons	12	108	poly:WS	36	3	2	0	0	21	1	6	39
bakery	10	90	poly:SW	73	0	0	0	0	2	1	2	12
banana	11	135	poly:WS	90	2	8	0	0	8	0	2	25
bandit	4	40	poly:SW	34	0	0	0	0	0	0	0	6
beeches	12	112	poly:SW	79	0	0	0	0	1	1	1	30
blacksmith	14	126	poly:SW	111	0	0	0	0	6	0	3	6
camera	11	103	poly:SW	87	0	0	0	0	3	1	1	11
canal	6	54	poly:WS	30	0	0	0	0	5	0	3	16
canoes	7	63	poly:WS	33	0	0	0	0	3	3	10	14
caravan	10	90	poly:SW	84	1	0	0	0	0	2	0	3
caravan	10	90	nasal:L	84	1	0	0	0	0	2	0	3
carved	16	188	ddel	33	0	2	0	0	2	2	2	147
cattle	17	161	poly:SW	141	0	0	0	0	1	0	3	16
cavalry	7	63	poly:SW	46	0	0	0	0	4	10	1	2
chapel	10	126	poly:SW	124	0	0	0	0	1	0	0	1
cobbled	10	126	ddel	91	0	8	0	13	0	1	0	13
coconut	6	54	poly:SW	45	0	0	0	0	0	0	1	8
collapsed	3	27	poly:WS	6	0	7	0	0	1	0	0	13
concealed	12	148	poly:WS	104	0	1	0	0	11	1	8	23
crane	10	94	nasal:L	20	0	0	0	0	0	1	0	73
crevasse	4	72	poly:WS	42	0	0	0	0	2	0	5	23
crevice	3	27	poly:SW	10	0	0	0	0	0	4	3	10
crocodiles	4	36	poly:SW	18	0	6	0	0	1	11	0	0
crossing	14	162	poly:SW	128	0	0	0	0	3	0	6	25
desert	7	63	poly:SW	50	0	0	3	0	0	1	3	6
diamond	15	175	ddel	158	1	0	5	0	0	3	0	8
disused	12	112	ddel	36	3	28	0	0	0	4	0	41
elephants	4	36	poly:SW	26	0	5	0	0	0	2	2	1
fallen	15	175	poly:SW	55	0	0	0	0	5	1	4	110
fallen	15	175	nasal:LV	55	0	0	0	0	5	1	4	110
farmed	6	54	ddel	14	0	10	0	0	4	0	2	24
flamingoes	7	63	poly:WS	54	0	2	0	0	4	0	2	1
forest	17	157	poly:SW	110	15	0	4	0	5	2	0	21

Table D.1: Classification of responses to target stimulus: all data

target	items	N	dataset	1	1*	1.	1+	1'	'–'	typo	NW	RWR
gazelles	3	27	poly:WS	11	1	4	0	0	1	9	1	0
giraffes	6	54	poly:WS	26	4	3	0	0	2	14	0	5
gold	18	202	ddel	100	0	0	0	0	11	0	1	90
golden	4	40	nasal:L	20	0	0	0	0	5	0	2	13
granite	7	67	poly:SW	49	0	0	0	0	1	4	6	7
green	10	90	nasal:L	59	0	0	0	0	2	0	0	29
indian	10	94	nasal:V	55	1	0	0	0	3	0	3	32
iron	10	90	nasal:L	63	0	0	0	0	6	0	3	18
lagoon	10	90	poly:WS	49	0	0	0	0	13	0	11	17
lemon	12	144	poly:SW	56	0	0	0	0	13	0	6	69
lemon	12	144	nasal:V	56	0	0	0	0	13	0	6	69
level	14	162	poly:SW	149	0	0	0	0	0	0	2	11
limestone	10	90	nasal:V	75	0	0	1	0	5	0	6	3
lion	6	54	nasal:V	15	0	0	0	0	3	0	0	36
machete	15	171	poly:WS	143	4	1	0	0	5	10	6	2
monastery	8	72	poly:SW	41	3	0	0	0	2	17	1	8
monument	4	36	poly:SW	24	0	0	0	0	1	0	0	11
old	14	130	ddel	31	0	0	0	0	23	0	4	72
overgrown	14	162	nasal:V	104	0	0	0	0	12	0	8	38
overnight	3	27	poly:SW	24	0	0	1	0	0	0	0	2
owned	11	135	ddel	16	0	0	0	3	29	0	5	82
pebbled	11	103	ddel	5	0	14	0	8	12	1	4	59
pelicans	3	27	poly:SW	16	0	7	0	0	1	0	1	2
picket	7	63	poly:SW	31	0	0	1	0	0	0	2	29
pillars	7	63	poly:SW	32	0	0	0	0	0	6	1	24
pine	8	80	nasal:V	19	0	0	0	0	5	0	0	56
poisoned	12	108	poly:SW	3	0	93	0	0	1	5	0	6
poisoned	12	108	ddel	3	0	93	0	0	1	5	0	6
popular	3	27	poly:SW	6	0	0	0	0	7	0	2	12
ravine	28	296	poly:WS	175	5	0	0	0	16	12	6	82
remote	16	148	poly:WS	95	0	0	0	0	19	0	7	27
rocket	4	36	poly:SW	34	1	0	0	0	0	0	1	0
roman	13	117	nasal:L	48	0	0	0	0	11	0	4	54
round	3	27	ddel	5	0	0	0	0	0	0	1	21
saloon	10	90	poly:WS	32	1	0	1	0	11	0	6	39
saloon	10	90	nasal:L	32	1	0	1	0	11	0	6	39
savannah	4	36	poly:WS	24	0	0	0	0	4	4	0	4
saxon	22	242	nasal:L	111	0	0	0	0	12	3	17	99

Table D.1: Classification of responses to target stimulus: all data (contd)

target	items	N	dataset	1	1*	1.	1+	1'	'–'	typo	NW	RWR
settlement	4	36	poly:SW	25	0	0	2	0	0	0	1	8
seven	12	112	poly:SW	96	0	0	0	0	4	1	0	11
seven	12	112	nasal:L	96	0	0	0	0	4	1	0	11
shelter	7	63	poly:SW	56	1	0	1	0	1	0	3	1
submerged	3	27	ddel	11	0	2	0	0	9	0	2	3
swan	3	27	nasal:L	17	0	0	0	0	0	0	1	9
telephone	18	170	poly:SW	167	0	0	0	0	0	1	1	1
telephone	18	170	nasal:L	167	0	0	0	0	0	1	1	1
territory	4	40	poly:SW	–	–	–	–	–	–	–	–	–
totem	10	126	poly:SW	43	0	0	0	0	3	5	2	73
tourist	3	27	poly:SW	20	0	0	0	0	0	0	0	7
train	3	27	nasal:V	25	0	0	0	0	0	0	0	2
tribal	4	36	poly:SW	16	0	0	0	0	1	6	2	11
village	20	220	poly:SW	204	0	0	0	0	4	0	3	9
wagon	6	54	poly:SW	34	2	0	0	0	3	0	1	14
walled	3	27	ddel	1	0	13	0	0	1	0	0	12
waterfall	14	130	poly:SW	119	1	0	0	0	1	0	2	7
waterhole	6	54	poly:SW	33	0	0	0	0	3	0	3	15

Table D.1: Classification of responses to target stimulus: all data (contd)

D.2 Classification of Real Word Responses, indicating how many RWRs matched to metrical structure, number of syllables, consonant onset and stressed vowel

Stimulus Word	N	Stress	Real Words	No of RWRs which match to: Metrical structure	Number Syllables	Stressed Vowel	Cons Onset
abandoned	27	WS	0	-	-	-	-
allotments	152	WS	7	0	0	3	0
apache	63	WS	2	0	0	2	0
baboons	108	WS	39	8	22	6	16
bakery	90	SW	12	11	10	9	10
banana	135	WS	25	10	7	8†	8
bandit	40	SW	6	1	1	6	0
beeches	112	SW	30	27	27	22	1
blacksmith	126	SW	6	1	1	5	4
camera	103	SW	11	10	9*	4	10
canal	54	WS	16	0	4	5†	7
canoes	63	WS	14	7	6	2	9
caravan	90	SW	3	3	0	3†	3
cattle	161	SW	16	13	9	13†	14
cavalry	63	SW	2	2	2	2	2
chapel	126	SW	1	1	1	1	0
coconut	54	SW	8	3	0	0	7
collapsed	27	WS	13	0	0	10†	11
concealed	148	WS	23	6	9	11	11
crevasse	72	WS	23	5	13	15†	6
crevice	27	SW	10	1	2	1	4
crocodiles	36	SW	0	-	-	-	-
crossing	162	SW	25	16	13	19‡	11
desert	63	SW	6	0	6	0	6
elephants	36	SW	1	1	0	0	0
fallen	175	SW	110	54	54	84‡	79
flamingoes	63	WS	1	0	0	0	1
forest	157	SW	21	9	8	3	9
gazelles	27	WS	0	-	-	-	-
giraffes	54	WS	5	2	2	1	0
granite	67	SW	7	7	7	7	2

†assuming [a] = [ɑ]; ‡assuming [ɒ] = [ɔ];
* if treated as a two syllable word.

Table D.2: a) Number of Real Word Responses to Polysyllabic stimuli which match to metrical stress, number of syllables, stressed vowel or consonant onset

Stimulus Word	N	Stress	Real Words	No of RWRs which match to: Metrical structure	Number Syllables	Stressed Vowel	Cons Onset
lagoon	90	WS	17	2	8	2	4
lemon	144	SW	69	37	38	21	25
level	162	SW	11	10	10	6	5
machete	171	WS	2	0	2	2	0
monastery	72	SW	8	8	8**	6	6
monument	36	SW	11	11	3	4‡	1
overnight	27	SW	2	1	1	1	1
pelicans	27	SW	2	1	1	2	0
picket	63	SW	29	29	29	28	1
pillars	63	SW	24	22	24	9	10
poisoned	108	SW	6	6	6	0	3
popular	27	SW	12	6	2	8	11
ravine	296	WS	82	45	74	25	46
remote	148	WS	27	14	21	17	3
rocket	36	SW	0	-	-	-	-
saloon	90	WS	39	4	19	2	31
savannah	36	WS	4	4	4	3†	2
settlement	36	SW	8	8	7	7	8
seven	112	SW	11	3	6	1	10
shelter	63	SW	1	1	1	1	0
telephone	170	SW	1	1	0	1	1
territory	40	SW	n/a	n/a	n/a	n/a	n/a
totem	126	SW	73	70	73	63	60
tourist	27	SW	7	2	2	1	0
tribal	36	SW	11	10	7	9	8
village	220	SW	9	7	7	2	2
wagon	54	SW	14	13	14	13	1
waterfall	130	SW	7	6	1	7	7
waterhole	54	SW	15	12	1	10‡	8

†assuming [a] = [ɑ]; ‡assuming [ɒ] = [ɔ];
* if treated as a two syllable word; ** if treated as a three syllable word.

Table D.2: a) Number of Real Word Responses to polysyllabic stimuli which match to metrical stress, number of syllables, stressed vowel or consonant onset (contd)

Stimulus Word	N	Dataset	Real Words	No of RWRs which match to: Metrical structure	Number Syllables	Stressed Vowel	Cons Onset
carved	188	ddel	147	145	145	110†	141
cobbled	126	ddel	13	13	13	0	12
diamond	175	ddel	8	6	7	6	5
disused	112	ddel	41	41§	41	41§	28
farmed	54	ddel	24	17	17	15†	14
gold	202	ddel	90	83	83	75	51
old	130	ddel	72	67	67	38	24
owned	135	ddel	82	72	72	23	32
pebbled	103	ddel	59	57	57	13	9
poisoned	108	ddel	6	6	6	0	3
round	27	ddel	21	21	21	7	10
submerged	27	ddel	3	0	0	2	0
walled	27	ddel	12	12	12	3	2
caravan	90	lab	3	3	0	3†	3
crane	94	lab	73	69	69	49	9
fallen	175	both	110	54	54	84‡	79
golden	40	lab	13	9	9	8	0
green	90	lab	29	19	19	16	5
indian	94	vel	32	23	31*	8	29
iron	90	lab	18	10	10	6	14
lemon	144	vel	69	37	38	21	25
limestone	90	vel	3	0	3	0	0
lion	54	vel	36	2	2	26	18
overgrown	162	vel	38	6	15	28	35
pine	80	vel	56	55	55	18	39
roman	117	lab	54	30	31	23	10
saloon	90	lab	39	4	19	2	31
saxon	242	lab	99	78	78	91	10
seven	112	lab	11	3	6	1	10
swan	27	lab	9	7	7	2	9
telephone	170	lab	1	1	0	1	1
train	27	vel	2	2	2	2	0

†assuming [a] = [ɑ]; ‡assuming [ɒ] = [ɔ];
* if treated as a two syllable word; §if treated as SW, with stressed vowel taken as [ɪ].

Table D.2: b) Number of Real Word Responses to stimuli from /d/-deletion and nasal assimilation dataset which match to metrical stress, number of syllables, stressed vowel or consonant onset